NOTE TO USERS

uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Jean-François Lécuyer
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Mathematics)
GRADE / DEGREE

Department of Mathematics and Statistics
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Comparison of Classification Trees and Logistic Regression to Model the Severity of Collisions Involving Elderly Drivers in Canada

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. G. Ivanoff
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. A. Alvo

Dr. S. Sinha

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

COMPARISON OF CLASSIFICATION TREES AND LOGISTIC REGRESSION TO MODEL THE SEVERITY OF COLLISIONS INVOLVING ELDERLY DRIVERS IN CANADA

Jean-François Lécuyer

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
for the degree of
Master of Science in Mathematics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Your file Votre référence
ISBN: 978-0-494-48470-8
Our file Notre référence
ISBN: 978-0-494-48470-8

Abstract

The number of drivers aged 65 years and older in Canada and the proportion of the population these drivers represent have been increasing for many years and will continue to do so in years to come. This increase in the number of elderly drivers could possibly lead to an increase in the numbers of fatalities, serious injuries and collisions involving drivers of this age group[1]. In order to find ways to reduce the number of collisions involving elderly drivers, and in particular the number of fatalities among the victims of collisions involving drivers aged 65 years and older, the relationship between the characteristics of these collisions and their severity was modeled using both classification trees and logistic regression. In this thesis, we explain the theory behind classification trees and logistic regression before analyzing the data. Both techniques are also compared based on the results of the analysis. In particular, we have validated the classification trees with the more rigorous logistic regression analysis. Consequently, the non-statistician can use the visually appealing trees with confidence.

Acknowledgements

I would like to thank Transport Canada for giving me the permission to use their data and particularly Aline Chouinard for her help. I would also like to thank my two supervisors, Gail Ivanoff and Mahmoud Zarepour, for their help.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The number of drivers aged 65 years and older in Canada and the proportion of the population these drivers represent have been increasing for many years and will continue to do so for the years to come. This increase in the number of drivers aged 65 years and older could possibly lead to an increase in the numbers of fatalities, serious injuries and collisions involving drivers of this age group[1].

To find ways to reduce the number of collisions involving drivers aged 65 years and older, and in particular the number of fatalities among the victims of collisions involving drivers aged 65 years and older, we need to know which characteristics of the collisions involving drivers of this age group have an effect on the severity of these collisions.

1.1 Purpose of the thesis

The primary purpose of this thesis is to answer the following question: Which characteristic of the collisions involving drivers aged 65 years and older have an effect on the severity of these collisions?

To model this relationship, two techniques were used: classification trees and logistic regression. Classification trees were used as they provide an easy to understand visual representation of the relationship between the response variable and the

predictor variables. Logistic regression was used as it provides a rigorous statistical analysis for categorical data. In particular, logistic regression provides us with odds ratios for each effect in the model, with their associated confidence intervals and the p-values to test the significance of each effect.

In this thesis, we will explain the theory behind classification trees and logistic regression before analyzing the data. Both techniques will also be compared based on the results of the analysis.

1.2 Review of the literature

Many studies have been conducted in the past to model the severity of motor vehicle collisions. However, to our knowledge, few of them were exclusively about collisions involving elderly drivers. In particular, two studies used logistic regression to model the relationship between the severity of collisions involving drivers aged 65 years and older and the different characteristics of these collisions.

Zhang, Lindsay, Clarke, Robbins and Mao (1999)[2] studied the factors having an effect on the severity of the collisions involving drivers aged 65 years and older in Ontario. The study was limited to collisions involving at least one driver aged 65 years and older which occurred between 1988 and 1993 and included only the collisions involving passenger cars, minivans and light trucks where only drivers and passengers were involved. Collisions involving other types of vehicles or other types of road users were excluded from the analysis. 34,927 collisions are included in this study, including 697 fatal collision. Multivariate logistic regression was used to model the risk of fatal collisions, major-injury collisions or minor-injury collisions compared to minimal-injury collisions. The study showed that the following factors significantly increased the risk of fatal collisions compared to minimal-injury collisions: age, sex, failing to yield right-of-way/disobeying traffic signs, non-use of seat belts, ejection from the vehicle, intersection without traffic controls, roads with higher speed limits, snowy weather, head-on collisions, two-vehicle turning collisions, overtaking and changing

lanes.

Dissanayake and Lu (2001)[3] studied the factors affecting the severity of fixed object-passenger car collisions where the driver was 65 years and older which occurred between 1994 and 1996 in Florida. Logistic regression models using either the severity of the collision or the severity of the injuries of the driver as response variable were fitted. The sample sizes for the two models were respectively 7,637 and 7,371. The following variables were found to affect the severity: travel speed, restraint device usage, point of impact, use of alcohol and drugs, personal condition, gender, whether the driver is at fault, urban/rural nature of the road and grade/curve existence at the collision location.

Other studies, although not exclusively about drivers aged 65 year and older, provided interesting results. Hill and Boyle (2005)[4] studied the risk of severe injuries to drivers aged 35 years and older in collisions in the United States in 2000. This study used more than 114,000 observations. The study showed that the risk of severe injuries was higher among drivers aged between 55 and 74 years old and drivers 75 years and older than among other drivers. Moreover, female drivers aged 55 to 74 years old were shown to be more likely to be seriously injured than other drivers.

Bédard, Guyatt, Stones and Hirdes (2001)[5] used multivariate logistic regression to model the risk that a driver was killed in single-vehicle collisions with fixed objects in the United States among fatal collisions which occurred between 1975 and 1998. Among other things, they found that drivers aged between 65 and 79 years old were 2.33 times more likely to be fatally injured than drivers between 40 and 49 years old. Drivers aged 80 years and older were 4.98 times more likely to be fatally injured. 12,325 observations were included in the analysis.

Mercier, Shelley, Rimkus and Mercier (1997)[6] studied the effect of the age and gender of the occupants as predictors for the severity of the injuries of these occupants in head-on collisions which occurred on Iowa highways. 2,171 observations were included in the analysis, including 336 fatal injuries. The age of the person was

found to have a significant effect on the risk of severe or fatal injuries. Older occupants are more likely to suffer more severe injuries than younger occupants. The interactions between the age of the occupant and its position in the vehicle and between the age and the protection of this occupant were also shown to be significant.

1.3 Data and methodology

The data used in this thesis is obtained from the National Collision Database (NCDB) [7]. NCDB contains data on all motor vehicle collisions reported annually to Transport Canada by the 13 jurisdictions. NCDB contains one record for each person involved in a collision in Canada and for each record, 69 variables covering different aspects of the collision are collected. The variables included in this database represent characteristics of these persons, of their vehicles and of the collisions in which they were involved. NCDB covers the period from 2000 to 2005. The data used in this thesis was obtained directly or indirectly from the variables in this database.

The relationship between the severity of the collisions involving drivers aged 65 years and older in Canada and the different characteristics of these drivers, of their vehicles and of the collisions in which they were involved was studied. Note that the drivers of all types of vehicles were included in the analysis

The six years of data available were used. However, as Quebec, Alberta and Nunavut had a high percentage of missing values in NCDB for some important variables, these three jurisdictions were excluded from the analysis. Collisions for which the number of vehicles involved in the collision was unknown were also excluded, as this number is needed to define some new variables. Only one observation was kept for each collision, meaning that for collisions where more than one of the drivers involved was aged 65 years and older, the characteristics of the vehicle and of the driver are those of the older driver.

To model the relationship between the severity of the collisions and their characteristics, two techniques appropriate for categorical data analysis were used: clas-

sification trees and logistic regression. The softwares S-Plus and SAS were used to perform the analysis. Classification trees were grown using the library *rpart* of S-Plus. The logistic regression models were fitted using PROC LOGISTIC in SAS with forward selection. Based on the fitted logistic regression models, odds ratios with their associated 95% confidence intervals were also obtained for each effect.

Two models were built using each technique. The first model uses the variable "Severity of collision" from NCDB as response variable. It models the risk of fatal collisions compared to non-fatal injury collisions.

Note that the variable "Severity of collision" is divided in 3 categories: "Collision producing at least one fatality", "Collision producing non-fatal injury" and "Collision involving property damage only". The severity of a collision is defined based on the severity of the worst injury produced by the collision. Thus, a fatal collision is a collision where there has been at least one fatality, a non-fatal injury collision is a collision where there has been at least one injury but no fatal injury and a collision with property damages only is a collision where no one was injured. However, as previous studies were only taking into account collisions with injuries [2], collisions where no one was injured were excluded from the analysis and thus only the first 2 categories of the variable "Severity of collision" are used. To simplify the notation, these categories were renamed "Fatal collision" and "Non-fatal injury collision". A total of 78,893 collisions were thus included in the analysis for this model. A total of 40 predictor variables are used in this model.

First note that some variables from NCDB were excluded from the analysis, as they don't provide any meaningful information that could be used in such a statistical analysis. These variables are: "Collision case number", "Collision identification number", "Police detachment/Region code", "Scene attended", "Vehicle identification number", "Vehicle sequence number", "Person sequence number" and "Special study". The variables "Province", "Day of the month", "Hit and run flag", "Road material", "Road condition", "Emergency use", "Province of driver's license", "Li-

cense status", and "Pedestrian action" were also excluded from the analysis as we think they don't affect the severity of the collisions.

The variables "Number of persons killed", "Number of persons injured", "Medical treatment required" and "Vehicle damage severity" were also excluded because they represent consequences of the collisions and not factors in the collision. The variable "Person position" was excluded as only the characteristics of the drivers are of interest in this thesis. The variables "Safety device used" and "Vehicle year" were not used directly in the analysis, but it was used to create other variables representing respectively the restraint use of the occupants and the vehicle age. The variable "Occupant ejection from vehicle" was excluded from the analysis because it is thought to be highly correlated with restraint use.

Moreover, some variables had a large percentage of missing values and were also excluded from the analysis. These variables are: "Artificial light condition", "Road classification II", "Road classification III", "Number of occupants in vehicle", "Vehicle use", "Trailer type", "Use of vehicle headlights", "Vehicle speed", "Vehicle event 2", "Vehicle event 3", "Prioritization of contributing factors", "Dangerous goods class", "Load status of commercial vehicles", "Years licensed in jurisdiction", "Ejection location", "Proper usage of safety device", "Blood alcohol concentration" and "Air bag deployment". Note that even though the variable "Vehicle speed" was excluded, two variables to be defined later take into account the speed of the vehicle: "Posted speed limit" and "Driving too fast for conditions". Likewise, the variable "Driver was under the influence of alcohol" will replace "Blood alcohol concentration".

Therefore, the following predictor variables, taken directly from NCDB, are used in this model (the categories for these variables appear between parenthesis, if necessary):

- Year of collision (2000, ..., 2005)
- Month of collision (January, ..., December)
- Day of the week (Monday to Friday, Saturday or Sunday)

- Hour of collision (0:00-2:59.,. . . , 21:00-23:59)
- Number of vehicles involved
- Collision configuration (Vehicle hit an object, Other single vehicle configuration, Rear-end collision, Two vehicle, same direction of travel configuration other than rear-end collision, Head-on collision, Right angle collision, Two vehicle, different direction configuration other than head-on or right angle collision)
- Roadway configuration (Non-intersection, At an intersection of at least two public roadways, Intersection with parking lot entrance/exit, private driveway or laneway, Other (bridge, overpass, tunnel, etc.))
- Weather condition (Clear and sunny, overcast or cloudy but no precipitation, Raining, Snowing, freezing rain, sleet, hail, Visibility limitation, Strong wind)
- Light condition (Daylight, Dawn or dusk, Darkness)
- Road classification I (Urban, Rural)
- Road surface (Dry or normal, Wet, Snow/Slush/Wet snow/Icy, Other (sand, gravel, dirt, oil etc.)
- Road alignment (Straight, Curved)
- Traffic control (Traffic lights, Sign, Other, No control present)
- Posted speed limit (50 km/h or less, 60 to 90 km/h, 100 km/h or more)
- Vehicle type (Passenger car, Passenger van or SUV, Light truck, Heavy truck, Bus, Motorcycle, Other)
- Vehicle manoeuvre (Going straight ahead, Turning left, Turning right, Making U-turn, Changing lanes, Reversing, Overtaking or passing, Slowing or stopping in traffic, Other)
- First impact location (Front, Roof, Rear, Left side, Right side, Other)
- Vehicle event 1 (Non-collision event, Hit moving object, Hit non-moving object)
- Driver sex (Male, Female)
- Driver age.

Note that for some of these variables, categories have been grouped together. Note also that the variables "Driver sex" and "Driver age" represent respectively the values of the variables "Person sex" and "Person age" for the driver.

Some other predictor variables were created, based on the variables in NCDB. In particular, NCDB contains 4 fields called "Contributing factor 1" to "Contributing factor 4" providing up to 4 factors which contributed to the collision. However, the data can't be used directly in the form it is given in NCDB. These 4 fields were thus recoded as 16 binary variables representing the presence or absence of each of 16 possible factors as a contributing factor to the collision. The 16 new predictor variables created are:

- Driver was fatigued or fell asleep
- Driver was under the influence of alcohol
- Driver was under the influence of drugs
- Other driver condition
- Following too closely
- Distraction, inattentiveness
- Driving too fast for conditions
- Improper turning or passing
- Failing to yield right-of-way
- Disobeying traffic control device or traffic officer
- Driving on wrong side of the road
- Lost control
- Other driver action
- Vehicle condition
- Environmental condition
- Other contributing factor.

The following 4 variables were also created to complete the list of predictor variables for this model:

- Vehicle age
- Number of unrestrained victims in the collision involving the studied driver
- Number of restrained victims in the collision involving the studied driver
- Number of victims who were not occupants of a vehicle (i.e. pedestrians, cyclists) in the collision involving the studied driver.

The variable "Vehicle age" was included in this model as we think it may have an effect on the severity of the collisions. It is obtained by subtracting the value of the variable "Vehicle model year" from the value of the variable "Year of collision". "Vehicle age" is set to 0 if this value happen to be smaller than 0. It can happen since new models for a given year are usually on the market the year before.

The variables related to the occupant's restraint status are defined based on the variable "Safety device used". These variables have been created as the restraint status of the occupants has been shown to affect the severity of the collisions in previous studies [2, 3]. A victim has been defined as any person who was injured in the collision. Restrained occupants include occupants wearing a lap and shoulder belt, a lap belt only and those in a child seat whereas unrestrained occupants are occupants not wearing a seat belt or wearing a shoulder belt only. Non-occupants are the persons who are not in one of the vehicles. All the victims in the collision involving the studied driver are thus counted in the variable corresponding to their restraint status. As it is impossible to determine from the data which driver was responsible in the collision, a weight of $\frac{1}{w}$ was assigned to each victim in the crash, where w is the number of vehicles involved in the collision. Thus, in a collision involving 2 vehicles, each victim would be assigned a weight of 0.5 and if there were 3 unrestrained victims in this collision, the variable "Number of unrestrained victims in the collision involving the studied driver" would have a value equal to 1.5. In a collision involving 3 vehicles, each victim would have a weight of $\frac{1}{3}$ and so on.

The second model uses the severity of the injuries to the driver, represented by the variable "Medical treatment required", as response variable. It models the risk of fatal injuries to the driver compared to non-fatal injuries. The variable "Medical treatment required" is divided in 7 categories: "No injury", "Minimal", "Minor", "Major", "Fatal", "Death or injury due to natural cause" and "Injured, extent unknown". However, drivers who were not injured or who suffered death or injuries due to a natural cause were excluded from the analysis as past studies didn't include them [2]. The remaining drivers are grouped such that the variable contains only two categories: "Fatally injured" and "Non-fatally injured".

There are 38 predictor variables used in this model. The first 37 variables are the same for both models. Only the predictor variables related to the restraint status are different. The three variables related to the restraint status in the first model are replaced by the variable (the categories for this variable are between parenthesis):

- Safety device used (Restrained, Unrestrained).

This variable represents the restraint status of the driver.

1.4 Organization of the thesis

This thesis is divided into 7 chapters. Chapter 2 covers the theory behind classification trees. In chapter 3, the data is analyzed using classification trees. Chapter 4 covers the theory behind logistic regression. In chapter 5, we analyze the data using logistic regression. The results are discussed in chapter 6. Chapter 7 will conclude the thesis.

Chapter 2

Classification trees

Classification and regression trees are methods used to predict the value of a given response variable based on the values of a number of predictor variables. The predicted values can either be categorical or continuous. Classification trees are used to predict the class of an observation if the response variable is categorical. Regression trees are used to predict the value of a response variable of the continuous type. The predictor variables can be either continuous, discrete, categorical or a combination of these three. These methods were developed somewhat recently, since their implementation was made possible only by the arrival of the computer. Before that, the computations would have been too tedious.

CART is the most well-known software for classification and regression trees. However, implementations exists in many other softwares, in particular in S-PLUS. In this case, classification and regression trees are implemented using either the *tree* or *rpart* functions.

Most of the theory presented in chapter 2 is based on Breiman, Friedman, Olshen and Stone (1984) [8]. Note that some material is also based on Therneau and Atkinson (1997)[9].

2.1 Growing a classification tree

The problem addressed in this thesis uses a categorial variable as predictor variable since we need to predict the severity of the collisions or of the injuries sustained by the road users, two categorical variables. Therefore, classification trees must be used. To grow a classification tree, we use a data set containing observations for which the value of a number of predictor variables is known. The value of the response variable also needs to be known for each observation, that is the classification of the observations according to this variable. The set formed by the response variable and the predictor variables is called the training set or training sample.

We first explain the method of growing a classification tree. Suppose we have a set of observations $\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})^T$ where $\mathbf{z_i} = (y_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n$. That is, $\mathbf{z}$ is represented by the matrix

$$\mathbf{z} = \begin{bmatrix} y_1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & \ldots & x_{nk} \end{bmatrix}$$

Y represents the response variable and $X_1, \ldots, X_k$ represent the predictor variables. This set contains all the observations and is called the root node of the tree. We need to find a split based on one of the predictor variables such that the reduction in the total impurity of the child nodes compared to the impurity of the root node obtained with this split is maximized. Since classification and regression trees are binary trees, it is important to note that each split must divide the observations between exactly two child nodes. Note that the child nodes of a given node t are called nodes $2t$ and $2t + 1$.

The impurity of a node is a measure of the homogeneity of the observations in that node. We want to obtain nodes which are as homogeneous as possible since a tree containing more homogeneous nodes results in lower error in the classification produced by this tree, and thus the tree is better. If for a given node, a significant

number of observations are classified in a different class than the observations in the most likely class, the impurity of the node is large. The maximum impurity should happen when each class is equally likely. On the contrary, if all the observations are classified in the same category, the impurity of the node should be 0. The node is then said to be pure. Several measures exist to assess the impurity. We now take a detailed look at some possible impurity criteria.

A plausible choice for a measure of node impurity would be the probability of misclassification of this node. The misclassified observations at a given node are the ones which are not in the most likely class among the observations in that node.

Define $P_Y(j) = P(Y = j), j = 1, \ldots, J$ to be the probabilities of an observation being in each of the J classes of the response variable. We call these prior class probabilities. These probabilities can be obtained by three ways:

1. If the sample is known to be representative of the population from which it is chosen, the empirical prior probabilities based on this sample are used as an estimate of the prior probabilities.

2. If the researcher has previous knowledge of the usual distribution of the observations between the classes, the prior probabilities can be chosen by the researcher

3. If no knowledge of the prior probabilities is available, they are usually taken to be equal. That is, $P_Y(j) = \frac{1}{j}, j = 1, \ldots, J$.

Let A represent the node in which the current observation lies. Then, define the conditional probability that an observation is in class j given that it is in node t as $P_{Y|A}(Y = j \mid \text{A=t}) = P_{Y|A}(j \mid t)$. As the prior probabilities of each class affects $P_{Y|A}(j \mid t)$, they must be chosen carefully. Note that in the analysis that will be conducted later, the empirical prior probabilities will be used.

Next, let $P_{Y,A}(j, t)$ be the probability that an observation is in node t and in class

j. This probability can be written as

$$P_{Y,A}(j,t) = P_{A|Y}(t \mid j) \times P_Y(j) = \frac{n_{jt}}{n_j} \times P_Y(j),$$

where n_{jt} is the number of observations in class j and node t and n_j is the total number of observations in class j.

If $P_A(t) = P(A = t)$ represents the probability that an observation is in node t, then

$$P_A(t) = \sum_j \frac{n_{jt}}{n_j} \times P_Y(j).$$

We can now define the probability that an observation is in class j given that the observation is in node t as

$$\begin{aligned} P_{Y|A}(Y = j \mid \text{A=t}) &= P_{Y|A}(j \mid t) \\ &= \frac{P_{Y,A}(j,t)}{P_A(t)} \\ &= \frac{P_{Y,A}(j,t)}{\sum_j P_{Y,A}(j,t)} \\ &= \frac{\frac{n_{jt}}{n_j} \times P_Y(j)}{\sum_j \frac{n_{jt}}{n_j} \times P_Y(j)}, \end{aligned}$$

Note that if $P_Y(j) = \frac{n_j}{n}$, where n is the number of observations in our data set, that is, the prior class probabilities are the empirical probabilities, we get:

$$P_{Y,A}(j,t) = \frac{n_{jt}}{n_j} \times \frac{n_j}{n} = \frac{n_{jt}}{n}.$$

Moreover, in this case, $P_A(t)$ becomes

$$P_A(t) = \sum_j \frac{n_{jt}}{n} = \frac{n_t}{n}.$$

Finally, $P_{Y|A}(j \mid t)$ becomes simply

$$P_{Y|A}(Y = j \mid \text{A=t}) = \frac{n_{jt}}{n_t}.$$

Thus, when the empirical prior probabilities are used, $P_{Y|A}(j \mid t)$ is much easier to calculate.

Now, let the most likely class for a given node t be the class j satisfying

$$j = argmax_{l=1,\ldots,J} P_{Y|A}(l \mid t)$$

Thus, the probability of misclassification at node t is given by

$$PM_t = \sum_{k \neq j} P_{Y|A}(k \mid t) = 1 - P_{Y|A}(j \mid t). \tag{2.1.1}$$

However, even though the probability of misclassification looks to be a good choice of impurity criterion, it has some undesired properties.

To see why this is the case, we look at what happens in a problem where the response variable has 2 classes. In such a problem, the probability of misclassification for a given node t would be given by

$$\begin{aligned} PM_t &= 1 - \max(P_{Y|A}(1 \mid t), P_{Y|A}(2 \mid t)) \\ &= \min(P_{Y|A}(1 \mid t), P_{Y|A}(2 \mid t)), \end{aligned}$$

However, as the response variable contains only two classes, we can write $P_{Y|A}(2 \mid t) = 1 - P_{Y|A}(1 \mid t)$. Thus, it is possible to represent the probability of misclassification as a function of $P_{Y|A}(1 \mid t)$ only. Thus, the probability of misclassification for node t then becomes

$$PM_t = \min(P_{Y|A}(1 \mid t), 1 - P_{Y|A}(1 \mid t)).$$

The probability of misclassification increases linearly when $P_{Y|A}(1 \mid t)$ is between 0 and $\frac{1}{2}$ and decreases linearly when $P_{Y|A}(1 \mid t)$ is between $\frac{1}{2}$ and 1.

However, our goal is to obtain nodes as pure as possible. Thus, if we have two splits giving the same total probability of misclassification for the child nodes, we would like to use a split in which one of the child nodes is almost pure instead of a split giving more equal probabilities of misclassification for the two child nodes. However, if we used the probability of misclassification as the impurity criterion, the split to be used would be chosen randomly. We would thus like to find an impurity function which would reward splits producing purer nodes. Thus, we would like to

find a function which decreases faster than linearly when $P_{Y|A}(1 \mid t)$ is close to 0 or 1, as such a function would choose a split for which one of the child nodes is almost pure over a split where both nodes have more similar impurity measures.

Thus, for this particular two class problem, we would like to find an impurity function I depending on $P_{Y|A}(1 \mid t)$ to estimate the probability of misclassification, but satisfying this property. Otherwise, the function should have the same properties as the probability of misclassification.

Thus, the function should first be minimized and maximized at the same points as the probability of misclassification. That is, the minimum, 0, should happen when $P_{Y|A}(1 \mid t)$=0 or 1. The maximum should happen when $P_{Y|A}(1 \mid t) = \frac{1}{2}$, that is, I should satisfy $I'(\frac{1}{2}) = 0$ and $I''(\frac{1}{2}) \leq 0$. 0, $\frac{1}{2}$ and 1 should also be the only critical points of the function, meaning that the function I should be strictly increasing for $0 \leq P_{Y|A}(1 \mid t) \leq \frac{1}{2}$ and strictly decreasing for $\frac{1}{2} \leq P_{Y|A}(1 \mid t) \leq 1$. So, the function I should be a concave function and it should thus satisfy $I''(P_{Y|A}(1 \mid t)) \leq 0$ for all $0 \leq P_{Y|A}(1 \mid t) \leq 1$.

Finally, as the probability of misclassification is a symmetric function, the function I should also be symmetric. That is, we need $I(P_{Y|A}(1 \mid t)) = I(1 - P_{Y|A}(1 \mid t))$.

To summarize, the impurity function I chosen to replace the probability of misclassification should satisfy the following four conditions:

1. $I(0) = I(1) = 0$
2. $I'(\frac{1}{2}) = 0$ and $I''(\frac{1}{2}) \leq 0$
3. $I(P_{Y|A}(1 \mid t)) = I(1 - P_{Y|A}(1 \mid t))$
4. $I''(P_{Y|A}(1 \mid t)) \leq 0 \; \forall \; 0 \leq P_{Y|A}(1 \mid t) \leq 1$

One of the groups of functions having the properties above are the functions of the form

$$I(P_{Y|A}(1 \mid t)) = a + b \times P_{Y|A}(1 \mid t) + c \times [P_{Y|A}(1 \mid t)]^2.$$

Many groups of functions satisfy these properties, but the functions of this group have a very simple form. We must now find the values of a, b and c needed to satisfy Condition (1)-(4). Condition (1) says that I must satisfy

$$I(0) = a + b \times 0 + c \times 0^2 = 0$$

$$I(1) = a + b \times 1 + c \times 1^2 = 0$$

By the first equation, we find that $a = 0$. The second equation says that we must have $a + b + c = 0$. But $a = 0$, so we need to have $b + c = 0$ or $c = -b$. Thus, I should be of the form

$$I(P_{Y|A}(1 \mid t)) = b \times P_{Y|A}(1 \mid t) - b \times [P_{Y|A}(1 \mid t)]^2.$$

Without loss of generality, let $b = 1$. Thus, we obtain

$$\begin{aligned} I(P_{Y|A}(1 \mid t)) &= P_{Y|A}(1 \mid t) - [P_{Y|A}(1 \mid t)]^2 \\ &= P_{Y|A}(1 \mid t)[1 - P_{Y|A}(1 \mid t)]. \end{aligned}$$

Remember that $P_{Y|A}(2 \mid t) = 1 - P_{Y|A}(1 \mid t)$. Thus, I can be written as

$$I(P_{Y|A}(1 \mid t)) = P_{Y|A}(1 \mid t)P_{Y|A}(2 \mid t).$$

A function I of this form also satisfies the other three conditions.

The derivative of I is

$$I'(P_{Y|A}(1 \mid t)) = 1 - 2P_{Y|A}(1 \mid t).$$

Set $I' = 0$. Then,

$$1 - 2P_{Y|A}(1 \mid t) = 0 \text{ or } P_{Y|A}(1 \mid t) = \frac{1}{2}.$$

So, $\frac{1}{2}$ is a critical point of the function I. Since $I''(P_{Y|A}(1 \mid t)) = -2$, then $I''(\frac{1}{2}) = -2 < 0$ and $\frac{1}{2}$ is thus a maximum of I. Conditions (2) and (4) are thus satisfied. Condition (3) is also clearly satisfied as $P_{Y|A}(2 \mid t) = 1 - P_{Y|A}(1 \mid t)$ and thus

$$I(P_{Y|A}(1 \mid t)) = P_{Y|A}(1 \mid t)P_{Y|A}(2 \mid t)$$

$$= P_{Y|A}(2 \mid t)P_{Y|A}(1 \mid t)$$
$$= I(1 - P_{Y|A}(1 \mid t)).$$

Thus, the impurity function I, given by

$$I(P_{Y|A}(1 \mid t)) = P_{Y|A}(1 \mid t)P_{Y|A}(2 \mid t)$$

can be used to replace the probability of misclassification.

We would now like to generalize this impurity function for problems where the response variable has more than two classes, say J classes. To achieve this, we first need to generalize Conditions (1)-(4) for such problems.

In the two class problem, Condition (1) says that I must satisfy $I(0) = I(1) = 0$. That is, I must be 0 only when $P_{Y|A}(1 \mid t) = 0$ and $P_{Y|A}(2 \mid t) = 1$ or $P_{Y|A}(1 \mid t) = 1$ and $P_{Y|A}(2 \mid t) = 0$. That is, I is minimized when all the observations are classified in the same class. Thus, Condition (1) can be generalized to

$$I(1, 0, \ldots, 0) = I(0, 1, 0, \ldots, 0) = \ldots = I(0, \ldots, 0, 1) = 0$$

Condition (2) for the two-class problem says $I(\frac{1}{2})$ should be the maximum of the function I. Thus, Condition (2) generalizes to saying that I should be maximized when each class is equally likely, or

$$I\left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right) = maximum.$$

Condition (3) says that I must be a symmetric function. That is, for any possible set of class probabilities $P_{Y|A}(j \mid t), j = 1, \ldots, J$, any permutation of the J class probabilities should give the same value for I.

Finally, in the two-class problem, Condition (4) says that the second derivative of $I(P_{Y|A}(1 \mid t))$ must be smaller than 0. That is, the function must be concave. We say that a function f is strictly concave if the Hessian of f is negative-definite.

Remember that the Hessian matrix of a function f at the point $\mathbf{x_0} = (x_1, \ldots, x_n)$

is defined as

$$Df(\mathbf{x_0}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

the matrix of the second partial derivatives of f at $\mathbf{x_0}$. Moreover, define $(Hf(\mathbf{x_0}))(h_1, \ldots, h_n)$ by

$$(Hf(\mathbf{x_0}))(h_1, \ldots, h_n) = \frac{1}{2} \begin{bmatrix} h_1 & \ldots & h_n \end{bmatrix} Df(\mathbf{x_0}) \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}.$$

For any convex set S, we say that the Hessian of f at the point $\mathbf{x_0}$, $Df(\mathbf{x_0})$, is negative-definite over S if $(Hf(\mathbf{x_0}))(h_1, \ldots, h_n) = 0$ if and only if $h_1 = \ldots = h_n = 0$ and $(Hf(\mathbf{x_0}))(h_1, \ldots, h_n) < 0$ otherwise, where $h_1, \ldots, h_n \in S$ [10]. Remember that a convex set is a set such that the line connecting any two points in the set lies in the set. Thus, the following four conditions must be satisfied in general by the impurity criterion I:

1. $I(1, 0, \ldots, 0) = I(0, 1, \ldots, 0) = \cdots = I(0, \ldots, 0, 1) = 0$
2. $I(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}) = maximum$
3. I is a symmetric function
4. I is a concave function or equivalently $DI(\mathbf{x_0})$ is negative-definite over $[0, 1] \times \ldots \times [0, 1]$

The impurity function $I(P_{Y|A}(1 \mid t)) = P_{Y|A}(1 \mid t)(1 - P_{Y|A}(1 \mid t))$ is generalized by the Gini criterion, given by

$$G_t = \sum_{k \neq j} P_{Y|A}(j \mid t) P_{Y|A}(k \mid t) = 1 - \sum_k P_{Y|A}(k \mid t)^2$$

since $\sum_{k \neq j} P_{Y|A}(j \mid t) P_{Y|A}(k \mid t) = (\sum_j P_{Y|A}(j \mid t))^2 - \sum_k P_{Y|A}(k \mid t)^2 = 1 - \sum_k P_{Y|A}(k \mid t)^2$ because $\sum_j P_{Y|A}(j \mid t) = 1$. The Gini criterion satisfies Condition (1)-(4).

Another criterion, the deviance, also satisfies these conditions. The deviance is defined as

$$D_t = -\sum_j P_{Y|A}(j \mid t) \log P_{Y|A}(j \mid t)$$

Even though the deviance was the first of the two criteria discovered, the Gini criterion is now the most frequently used in practice. We will show that these two criteria satisfy Condition (1)-(4) and thus that they can be used as impurity criteria.

Theorem 2.1.1 *The Gini criterion for node t,*

$$G_t = \sum_{k \neq j} P_{Y|A}(j \mid t) P_{Y|A}(k \mid t)$$

satisfies Condition (1)-(4).

Proof. It is straightforward to show that Condition (1) is satisfied, that is, $G_t(1, 0, \ldots, 0) = G_t(0, 1, 0, \cdots, 0) = \ldots = G_t(0, \ldots, 0, 1) = 0$. In each of these cases, one of the probabilities $P_{Y|A}(k \mid t), k = 1, \ldots, J$, is 1. All other probabilities are 0. Thus, none of the products $P_{Y|A}(j \mid t) P_{Y|A}(k \mid t)$, where $k \neq j$, can be different from 0, as only one of the two terms of the product can be different from 0. Thus, $G_t(1, 0, \ldots, 0) = G_t(0, 1, 0, \ldots, 0) = \cdots = G_t(0, \ldots, 0, 1) = 0$.

Now, we prove Condition (2). We want to show that

$$argmax_{0 \leq P_{Y|A}(j|t) \leq 1, j=1,\ldots,J} \; G_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(j \mid t)) = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)$$

To achieve this, we must first find the critical points of the function G_t. G_t can be written as a function of $J - 1$ variables only since $P_{Y|A}(1 \mid t) + \cdots + P_{Y|A}(J \mid t) = 1$ as this is a p.d.f. Thus,

$$P_{Y|A}(J \mid t) = 1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J - 1 \mid t)$$

It will be easier to work with such a function.

Now, we find the partial derivatives of G_t, given by

$$\frac{\partial G_t}{\partial P_{Y|A}(1 \mid t)} = -4P_{Y|A}(1 \mid t) - 2P_{Y|A}(2 \mid t) - \cdots - 2P_{Y|A}(J - 1 \mid t) + 2$$

$$\vdots$$

$$\frac{\partial G_t}{\partial P_{Y|A}(J-1 \mid t)} = -2P_{Y|A}(1 \mid t) - 2P_{Y|A}(2 \mid t) - \cdots - 4P_{Y|A}(J-1 \mid t) + 2$$

and equate them to 0.

Solving this system of equations, we find that these equations are all satisfied when $P_{Y|A}(1 \mid t) = \cdots = P_{Y|A}(J-1 \mid t) = \frac{1}{J}$ and thus $P_{Y|A}(J \mid t) = 1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t) = 1 - \frac{J-1}{J} = \frac{1}{J}$ too. Hence, $(\frac{1}{J}, \ldots, \frac{1}{J})$ is a critical point of G_t. However, the extremities of the region over which the probabilities are defined are also critical points, that is, $(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ are also critical points. However, for these latter points, $G_t = 0$ as shown before. On the contrary, for the point $(\frac{1}{J}, \ldots, \frac{1}{J})$, $G_t = \sum_{k \neq j} \frac{1}{J} \times \frac{1}{J} = \sum_{k \neq j} \frac{1}{J^2} = \frac{J(J-1)}{J^2} = \frac{J-1}{J}$, the maximum value among the critical points. Thus, G_t is maximized at the point $(\frac{1}{J}, \ldots, \frac{1}{J})$ and Condition (2) is satisfied.

Condition (3) is clearly true as the Gini criterion is a sum of all the possible products of the form $P(k \mid i)P(j \mid i)$ where $k \neq j$. Thus, when permuting the probabilities, we end up with the same terms in the sum and thus the same value for G_t. Hence, G_t is a symmetric function.

Finally, we prove Condition (4). That is, we want to show that G_t is a concave function. So, we must show that the Hessian of G_t is negative-definite.

The Hessian of G_t, written as a function of $J-1$ variables, is given by

$$DG_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)) = \begin{bmatrix} -4 & -2 & \ldots & -2 \\ \vdots & \vdots & \vdots & \vdots \\ -2 & -2 & \ldots & -4 \end{bmatrix}$$

Thus, HG_t is given by

$$\begin{aligned} &(HG_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)))(h_1, \ldots, h_{J-1}) \\ &= \frac{1}{2} \begin{bmatrix} h_1 & \ldots & h_{J-1} \end{bmatrix} \times \end{aligned}$$

$$DG_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)) \begin{bmatrix} h_1 \\ \ldots \\ h_{J-1} \end{bmatrix},$$

where $DG_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t))$ is as defined in the preceding equation.

Thus,

$$(HG_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)))(h_1, \ldots, h_{J-1}) = -2(\sum_{j,k} h_j h_k)$$

which takes negative values for all $(h_1, \ldots, h_{J-1})$ in the set $[0,1] \times \ldots \times [0,1]$ except when $(h_1, \ldots, h_{J-1}) = (0, \ldots, 0)$. Thus, the Hessian of G_t is negative definite over $[0,1] \times \ldots \times [0,1]$ and G_t is concave.

So, G_t satisfies Condition (1)-(4). □

We will now prove that the deviance also satisfies Condition (1)-(4).

Theorem 2.1.2 *The deviance of node t,*

$$D_t = -\sum_j P_{Y|A}(j \mid t) \times \log P_{Y|A}(j \mid t)$$

satisfies Condition (1)-(4).

Proof. It is straightforward to show that Condition (1) is satisfied, that is, $D_t(1,0,\ldots,0) = D_t(0,1,0,\ldots,0) = \cdots = D_t(0,\ldots,0,1) = 0$. First, note that $\forall\ j$, if $P_{Y|A}(j \mid t) = 0$, then $P_{Y|A}(j \mid t) \log P_{Y|A}(j \mid t) = 0 \times \log 0 = 0$ and if $P_{Y|A}(j \mid t) = 1$, then $P_{Y|A}(j \mid t) \log P_{Y|A}(j \mid t) = 1 \times \log 1 = 1 \times 0 = 0$. Thus, if $\forall\ j$, $P_{Y|A}(j \mid t)$ is either 0 or 1, then the terms of the sum are all 0 and thus, it is clear that $D_t(1,0,\ldots,0) = D_t(0,1,\ldots,0) = \cdots = D_t(0,\ldots,0,1) = 0$

Now, we prove Condition (2). Thus, we want to show that

$$argmax_{0 \le P_{Y|A}(j|t) \le 1, j=1,\ldots,J}\ D_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J \mid t)) = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)$$

To achieve this, we must first find the critical points of D_t. D_t can be written as a function of $J-1$ variables, as we have done for G_t. Let

$$P_{Y|A}(J \mid t) = 1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t).$$

Now, we find the partial derivatives of D_t, given by

$$\begin{aligned}
\frac{\partial D_t}{\partial P_{Y|A}(1 \mid t)} &= -\log(P_{Y|A}(1 \mid t)) + \log(1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t)) \\
&\vdots \\
\frac{\partial D_t}{\partial P_{Y|A}(J-1 \mid t)} &= -\log(P_{Y|A}(J-1 \mid t)) + \log(1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t))
\end{aligned}$$

and equate them to 0.

Solving this system of equations, we find that all equations are satisfied when $P_{Y|A}(1 \mid t) = \ldots = P_{Y|A}(J-1 \mid t) = 1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t) = \frac{1}{J}$. Thus $P_{Y|A}(J \mid t) = 1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t) = \frac{1}{J}$ too. Hence, $(\frac{1}{J}, \ldots, \frac{1}{J})$ is a critical point of D_t. However, the extremities of the region over which the probabilities are defined are also critical points. That is, $(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ are also critical points. However, for these latter points, $D_t = 0$ as shown before. On the contrary, for the point $(\frac{1}{J}, \ldots, \frac{1}{J})$, $D_t = -\sum_j \frac{1}{J} \log \frac{1}{J}$ which is greater than 0. Thus, D_t is maximized at the point $(\frac{1}{J}, \ldots, \frac{1}{J})$ and Condition (2) is satisfied.

Condition (3) is clearly true since permuting the probabilities doesn't change the terms included in the sum, so we end up with the same value for D_t. Hence, D_t is a symmetric function.

Finally, we prove Condition (4). That is, we want to show that D_t is a concave function. Thus, we must show that the Hessian of D_t is negative-definite over $[0,1] \times \ldots \times [0,1]$.

The Hessian of D_t, written as a function of $J-1$ variables, is given by the matrix with elements k_{ij} where

$$k_{ij} = \begin{cases} -\frac{1}{1 - P_{Y|A}(1|t) - \cdots - P_{Y|A}(J-1|t)} & \text{if } i \neq j \\ -\frac{1}{P_{Y|A}(j|t)} - \frac{1}{1 - P_{Y|A}(1|t) - \cdots - P_{Y|A}(J-1|t)} & \text{if } i = j \end{cases}$$

Thus, HD_t is given by

$$(HD_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)))(h_1, \ldots, h_{J-1})$$
$$= \frac{1}{2}\begin{bmatrix} h_1 & \ldots & h_{J-1} \end{bmatrix} DD_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t)) \begin{bmatrix} h_1 \\ \ldots \\ h_{J-1} \end{bmatrix}$$

where $DD_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J-1 \mid t))$ is as defined in the preceding equation.

Thus,

$$(HD_t(P_{Y|A}(1 \mid t), \ldots, P_{Y|A}(J \mid t)))(h_1, \ldots, h_{J-1})$$
$$= -\frac{1}{2}\left(\frac{h_1^2}{P_{Y|A}(1 \mid t)} + \cdots + \frac{h_{J-1}^2}{P_{Y|A}(J-1 \mid t)} + \frac{(h_1 + \cdots + h_{J-1})^2}{1 - P_{Y|A}(1 \mid t) - \cdots - P_{Y|A}(J-1 \mid t)}\right)$$

which is always negative for all $(h_1, \ldots, h_{J-1})$ except when $(h_1, \ldots, h_{J-1}) = (0, \ldots, 0)$, as the expression in the parenthesis is clearly positive. Thus, the Hessian of D_t is negative definite and so D_t is concave.

So, D_t satisfies Conditions (1)-(4). □

One of these impurity measures is chosen at the beginning to be used to grow the tree. As stated before, classification trees are binary trees, so we must find a split generated by a question of the form "Is $x_m < a$?" for some $m = 1, \ldots, k$ and some number a if x_m is numerical, or of the form "Is $x_m \in A$?" for some $m = 1, \ldots, k$ and some set of classes A if x_m is categorical (although, in general, it can be in the form of any question with a yes/no answer, but these are more difficult to implement on a computer program). If x_m is an ordered categorical variable, the set A must be such that the classes contained in it are consecutive based on the ordering of the categories.

For numerical variables, splits should be generated by questions of the form "Is $x_m < a$?" where $a \in R$. Any value $a \in R$ could be used as a split point. However, as an example, suppose 4 and 5 are two consecutive values in the training set for a particular problem when they are ordered. Then, the questions "Is $x_m < 4.3$?" and

"Is $x_m < 4.5$?" generate the same splits as no values are contained between 4.3 and 4.5. To make sure that the same split is not used twice, we usually choose as possible split points the midpoints between each pair of consecutive values present in the data set. Thus, in the current example, the question "Is $x_m < 4.5$?" would be used to generate the split sending the observations for which $x_m \leq 4$ to the left child node and those for which $x_m \geq 5$ to the right child node. Hence, there is a finite number of possible splits for numerical variables.

For categorical variables, the possible splits are the splits generated by questions of the form "Is $x_m \in A$?" where A can be any subset of the categories of x_m. For ordered categorical variable, the only difference is that the categories in A must be consecutive in the determined ordering of the categories.

For each observation, if the answer to the question is yes, the observation is sent to the left child node of t, and if the answer is no, it is sent to the right child node of t. For each variable x_m, we calculate the impurity of each child node and then the total impurity of the child nodes for each possible split and we find the split producing the greatest reduction in the impurity of the child nodes. The total impurity of the child nodes is the sum of the impurities of the child nodes weighted by the probability that an observation from the parent node is sent to each child node, respectively $\frac{P_A(2t)}{P_A(t)}$ and $\frac{P_A(2t+1)}{P_A(t)}$ for child nodes $2t$ and $2t+1$. In the case where the prior probabilities are the empirical ones, these probabilities become $\frac{\frac{n_{2t}}{n}}{\frac{n_t}{n}} = \frac{n_{2t}}{n_t}$ and $\frac{\frac{n_{2t+1}}{n}}{\frac{n_t}{n}} = \frac{n_{2t+1}}{n_t}$

This split is the best at that point in the tree. However, it is not guaranteed that the tree obtained when we finish splitting will be the best to classify the observations from the training set since a different choice of split at this point can lead to better splits further in the tree. That is, the reduction in impurity obtained with this choice of split might be the most significant at this level but can lead to smaller reductions in impurity further in the tree than if another split with a less significant reduction in impurity was chosen at this point. Given the split obtained, the observations are divided between the two child nodes depending on the answer to the splitting

question.

After the root node has been split, we are left with two child nodes. For each of these nodes, the same steps that were applied for the root node are repeated to find the way to split each of these nodes so as to maximize the reduction in impurity. The splitting of a given node is based on the observations assigned to that node based on the preceding split.

We continue to split the nodes obtained if

1. they are not pure (that is, the observations in the node are not all in the same class) or

2. a pre-determined threshold for the minimum number of observations needed in a node to continue splitting has not been reached.

When all the nodes in the tree cannot be split further for one of these two reasons, the splitting is stopped. The nodes at the bottom of the tree are called terminal nodes. The tree impurity is the weigthed sum of the impurities of these nodes.

Observations can then be assigned to a class for each terminal node. For a given terminal node, its class is the most likely class in that node based on the value of the response variable for all the observations. The most likely class is the class j for which $P_{Y|A}(j \mid t)$ is maximized for the observations in our data set. All the observations in the node are classified as being in that class.

2.2 Pruning the tree

When each node in the tree has been split until one of the two conditions above is satisfied, it is probable that there is some over fitting. In such a case, the tree is too specific to the data used to grow it. Hence, the tree would not give good results when classifying a set of new observations. To prevent such a situation from happening, pruning is used. Pruning means removing nodes, beginning at the bottom of the tree.

That is, once the full tree has been built using the method outlined in section 2.1, the idea is to remove nodes from the tree until the tree is optimal in some sense.

Trees are usually pruned using minimal cost-complexity pruning. We will look at minimal cost-complexity pruning in more details. Let T be a tree grown based on a set of observations $\mathbf{z}$ and let $\mid T \mid$ be the size of T, represented by the number of terminal nodes. Define the cost-complexity criterion of T for the training set $\mathbf{z}$ by

$$C_\lambda(\mathbf{z}, T) = C(\mathbf{z}, T) + \lambda \mid T \mid$$

Usually, $C(\mathbf{z}, T)$, the cost of the tree T, is taken to be the probability of misclassification in the case of classification trees and λ is a penalty for the size of the tree, called the complexity parameter. The value of λ will be defined later. In minimal cost-complexity pruning, the goal is to minimize the cost of the tree for general data sets.

Pruning works as follows. Suppose a tree T has already been grown based on a given sample. Define T_t as the branch of the tree T having node t as root node. Thus, T_t is a subtree of T and its cost can be calculated as for any tree T. Define also the cost of a node t alone, $C(\mathbf{z}, t)$, as the cost of the tree containing only node t. This cost is calculated based on the observations in node t only. Beginning with the tree T, if it contains some pairs of terminal nodes for which the sum of their total cost is equal to the cost of the node from which they were obtained, we remove these nodes. The tree obtained is called T_1. We begin the pruning process with T_1 as the original tree.

We first note that for any branch T_t of T_1, where t is not a terminal node,

$$C(\mathbf{z}, t) > C(\mathbf{z}, T_t)$$

as splitting nodes reduces or does not change the cost of a tree and the tree T_t is obtained by splitting the node t. Moreover, T_1 is defined such that the splits not reducing the cost are removed. This means that the cost of the tree containing only node t is greater than the cost of the tree T_t.

Moreover, for any branch T_t, the cost-complexity criterion of this branch must be less than or equal to the cost-complexity criterion of the tree consisting of node t only, and T_t is a better subtree than t, as long as

$$C_\lambda(\mathbf{z}, T_t) < C_\lambda(\mathbf{z}, t).$$

However, when $C_\lambda(\mathbf{z}, T_t)$ becomes equal to $C_\lambda(\mathbf{z}, t)$, the node t becomes a better subtree than T_t. Although the cost-complexity is the same at this point, the best tree is defined as the tree minimizing the cost-complexity criterion, and in case of equality, as the smaller tree in terms of nodes. Now, we must find the value of λ at which this becomes the case. For a given t, we denote this value by $\lambda(t)$ as the value of λ depends on the root node t of T_t. The branch T_t is better than the subtree consisting of the node t only when

$$\begin{aligned} & C_\lambda(\mathbf{z}, T_t) < C_\lambda(\mathbf{z}, t) \\ \Rightarrow \quad & C(\mathbf{z}, T_t) + \lambda(t) \mid T_t \mid < C(\mathbf{z}, t) + \lambda(t) \mid t \mid \\ \Rightarrow \quad & \lambda(t) \mid T_t \mid - \lambda(t) < C(\mathbf{z}, t) - C(\mathbf{z}, T_t) \\ \Rightarrow \quad & \lambda(t) < \tfrac{C(\mathbf{z},t) - C(\mathbf{z},T_t)}{\mid T_t \mid - 1}. \end{aligned}$$

Thus, we find that T_1 is a better tree when $\lambda_1 = 0 \leq \lambda(t) < \frac{C(\mathbf{z},t)-C(\mathbf{z},T_t)}{|T_t|-1}$, but $T_1 - T_t$ becomes a better tree when $\lambda(t) \geq \frac{C(\mathbf{z},t)-C(\mathbf{z},T_t)}{|T_t|-1}$.

So, for each non-terminal node $t_i, i = 1, \ldots$, we find the threshold $\lambda_1(t_i)$ at which $T_1 - T_t$ becomes a better tree than T_1. Among all non-terminal nodes, we find the node t_1 s.t. $t_1 = argmin_{t_i, i=1,\ldots} \lambda_1(t_i)$. t_1 is the first node such that $C_\lambda(\mathbf{z}, T_{t_i}) \geq C_\lambda(\mathbf{z}, t_i)$ for some t_i as λ increases. Then, we prune the branch T_{t_1} from the tree T_1 to obtain a subtree, denoted T_2, where $T_2 = T_1 - T_{t_1}$. The complexity parameter associated with this tree is $\lambda_2 = \lambda_1(t_1)$.

Based on T_2 we repeat the same operations. A new tree, T_3, is obtained, with associated complexity parameter λ_3. We continue this process until we have a tree containing only the root node of T_1.

Thus, we now have a sequence of subtrees of T, $T_1, T_2, \ldots$, with associated complexity parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Cross-validation is then used to estimate the cost of the tree T_i, $C(\mathbf{z}, T_i), i = 1, \ldots, k,$. Cross-validation works as follows. Suppose we want to estimate a value depending on a given population, such as the probability of misclassification in the case of trees. Denote this value by R. k-fold cross-validation can be used to do this.

k-fold cross-validation randomly divides a data set into k subsets of approximately equal size. A first estimate of R is obtained based on a tree grown from $k-1$ of the subsets. These subsets are used as the training set based on which the tree is grown whereas the remaining set serves as a test set from which the estimate of R is calculated. This is repeated k times, each time with a different subset used as the test set, and thus k estimates of R, denoted $R_1, \ldots, R_k$, are obtained. The average of these estimates,

$$\overline{R} = \frac{1}{k} \sum_{i=1}^{k} R_i,$$

is called the cross-validation estimate of R and usually gives a better estimate of than the one obtained by using the whole set as a test set.

Remember that the cost of a tree is the probability of misclassification of this tree. Remember also that the definition of the probability of misclassification for a given node is given by equation (2.1.1). Thus, the probability of misclassification for the whole tree, based on this equation, is given by

$$PM = \sum_{t} PM_t \times P_A(t) \tag{2.2.1}$$

where $P_A(t)$ is the probability that an observation is in node t. A plot of the cross-validation estimates of $C(\mathbf{z}, T_i)$ against the corresponding size or complexity parameter λ_i can then be drawn to get a better look at the relationship between the two values.

The value of i minimizing the cross-validation estimate of $C(\mathbf{z}, T_i)$ is then used to prune the tree. The tree associated with the chosen i is thus selected as the best

pruned tree. The tree obtained is usually a good estimate of the overall best tree in terms of cost-complexity.

As cross-validation selects randomly the k subsets it uses, it may be repeated a few times to check if the results are always the same. Once the best pruned tree has been obtained, observations can be assigned to classes as in section 2.1.

2.3 Obtain an estimate of the probability of misclassification of a tree

The probability of misclassification of a tree, given by equation (2.2.1) gives an indication of how well a tree classifies the observations of a given set.

When a set of observations is dropped into the tree, they are assigned to different classes according to which terminal node they end up in. Thus, an easy way to estimate the probability of misclassification of a tree is by dropping into the tree the observations used to grow it and calculate the probability of misclassification based on these observations. However, it is usually too optimistic an estimate of the true average probability of misclassification for more general data sets since the same data is used to grow the tree and to calculate the probability of misclassification. It is clear that a tree should better classify the data used to grow it than other data. There are two ways to obtain a better estimate of the true average probability of misclassification of a tree for a general set of observations.

The first way is to test it on a different set of observations for which the value of the response variable is known for all the observations, called a test set. If there is no other set on which we can test the data, we can divide the data set into two subsets before growing the tree, a training set and a test set. The training set is used to grow the tree. Then the observations of the test set are dropped into the tree and the probability of misclassification of the tree is estimated. Cross-validation, defined in section 2.2, is also another way to obtain a better estimate of the probability of

misclassification of the tree.

2.4 Missing values

Missing values for some variables in the data set can lead to problems when growing a tree or using a tree to classify observations.

We first look to what happens when we are growing a tree using a set of observations with missing values for some of the variables. Observations for which the value of the response variable is missing should clearly be removed from the data set used to grow the tree as we don't know the true class of these observations.

Observations where the values of some predictor variables are missing can still be used to grow the tree. Remember that at any point in growing the tree, we want to find the split producing the greatest reduction in tree impurity.

To achieve this, all possible splits at this node are tested to see which one gives the largest impurity reduction. Suppose some variable x_m has missing values for some observations. Then, the splits based on this variable are computed using only the observations for which the value for this variable is available. The same technique is used for all the variables with missing values. Once all possible splits for a given variable have been found, the split producing the greatest impurity reduction is chosen as before.

However, when we are trying to classify observations with missing values for some variables, it is possible that at some point in the tree, the value of the variable we need to split on is not available. This means that the observations in question could not be classified. We could completely remove such observations from the set of observation to classify. However, instead of removing these observations, we can handle them by using surrogate splits. Surrogate splits are used to classify the observations which cannot be assigned to a terminal node because the value of a variable used in some split is missing for this observation. Surrogate splits are defined as the splits producing the classification closest to the one obtained with the original split. The first surrogate

split is the split with the highest proportion of observations sent to the same node as by the best split (using the observations which can be classified using both splits, since if observations can't be classified by both splits, we can't compare where these observations are sent by the two splits).

Suppose we need to find surrogate splits for a given node t. Let s be the best split found for node t when the tree was originally grown. Let x_m be one of the predictor variables and let $s_{m1}, \ldots, s_{mp}$ be the set of all possible binary splits of node t based on the variable x_m. The splits are based on the observations without missing values for the variable on which the split is based. Suppose s_{mi} is one of the splits from this set. Let $n_j, j = 1, \ldots, J$, be the number of class j observations in the data set and let $P_Y(j) = P(Y = j)$ be the prior probability for class j (based on the training set). Also, define n_{jL}=number of class j observations sent to the left child node of t by both the splits s and s_{mi} and n_{jR}=number of class j observations sent to the right child node of t by both s and s_{mi}.

Let L_s be the event that an observation is sent to the left child node of t by the split s and $L_{s_{mi}}$ be the event that an observation is sent to the left child node of t by the split s_{mi}. Thus, the probability that any observation from the data set is sent to the left child node of t by both splits is

$$P(L_s \cap L_{s_{mi}}) = \sum_{j=1}^{J} P_Y(j) \times \frac{n_{jL}}{n_j},$$

that is, the sum of the proportions of observations from each class sent to the left child node of t weighted by the probability that an observation is in the given class.

Similarly, let R_s be the event that an observation is sent to the right child node of t by the split s and $R_{s_{mi}}$ be the event that an observation is sent to the right child node of t by the split s_{mi}. Thus, the probability that any observation from the data set is sent to the right child node of t by both splits is

$$P(R_s \cap R_{s_{mi}}) = \sum_{j=1}^{J} P_Y(j) \times \frac{n_{jR}}{n_j}.$$

Thus, the probability that the splits s and s_{mi} both send an observation already known to be in node t to its left child node is

$$\begin{aligned} P(L_s \cap L_{s_{mi}} \mid t) &= P(\text{splits } s \text{ and } s_{mi} \text{ both send the observation to the left} \\ &\quad \text{child node of } t \mid \text{observation is in node } t) \\ &= \frac{P(L_s \cap L_{s_{mi}})}{P_A(t)}. \end{aligned}$$

where $P_A(t)$ is the probability that an observation is in node t.

Similarly, we can define

$$\begin{aligned} P(R_s \cap R_{s_{mi}} \mid t) &= P(\text{splits } s \text{ and } s_{mi} \text{ both send the observation to the right} \\ &\quad \text{child node of } t \mid \text{observation is in node } t) \\ &= \frac{P(R_s \cap R_{s_{mi}})}{P_A(t)}. \end{aligned}$$

Thus, the proportion of observations from node t which are sent to the same child node by both the splits s and s_{mi}, denoted $P(s, s_{mi} \mid t)$, is the sum of these two probabilities:

$$P(s, s_{mi} \mid t) = P(L_s \cap L_{s_{mi}} \mid t) + P(R_s \cap R_{s_{mi}} \mid t).$$

For each possible split s_{mi}, $i = 1, \ldots, p$, based on the variable x_m, this probability is calculated. The surrogate split based on the variable x_m is the split based on x_m maximizing this probability. That is, we need to choose the split s_{ml} such that:

$$P(s, s_{ml} \mid t) = \max_{s_{mi}} P(s, s_{mi} \mid t), i = 1, \ldots, p.$$

This process is repeated for each predictor variable $x_m, m = 1, \ldots, k$, except the variable on which the split s is based. For each of the surrogate splits obtained, the predictive measure of association of the splits s and s_{ml} is then calculated. The predictive measure of association gives a measure of how well a split predicts another split and removes the splits that are not good enough.

Let $P(L_s \mid t)$ be the probability that an observation in node t is sent to the left child node of t by the split s. Then,

$$P(L_s \mid t) = \frac{P(L_s, t)}{P_A(t)} = \frac{P(L_s)}{P_A(t)}$$

since an observation sent to the left child node of t must already be in the node t. $P(L_s)$ is the probability that any observation is sent to the left child node of t by the split s.

Similarly, $P(R_s \mid t)$ is the probability that an observation in node t is sent to the right child node of t by the split s and is defined as

$$P(R_s \mid t) = \frac{P(R_s, t)}{P_A(t)} = \frac{P(R_s)}{P_A(t)}$$

since an observation sent to the right child node of t must already be in the node t. $P(R_s)$ is the probability that any observation in the tree is sent to the right child node of t by the split s.

Define a new split s' as follows. If $\max(P(L_s \mid t), P(R_s \mid t)) = P(L_s \mid t)$, the split s' sends all the observations in the node t to the left child node of t and no observations to the right child node of t. Otherwise, the split s' sends all the observations to the right child node of t. The probability that this split predicts correctly the split s is $\max(P(L_s \mid t), P(R_s \mid t))$ as the observations classified in the most likely child node of t by s are classified in the same child node by the split s'. However, the remaining observations are clearly classified in different child nodes by the splits s' and s. Thus, the probability of error of the split s' is $\min(P(L_s \mid t), P(R_s \mid t))$. The split s' is not a good split, particularly when $P(L_s \mid t)$ and $P(R_s \mid t)$ are almost equal, in which case the error probability is close to 50%. In such a case, the split s' is just slightly better than assigning observations at random to the child nodes

As the split s' is one of the worst possible splits, it would be natural to compare our possible surrogate splits to this split. We would in fact like to find a split much better than s'. Thus, for a given predictor variable x_m, define the predictive measure

of association of the splits s and s_{ml} as

$$A(s, s_{ml}) = \frac{\min(P(L_s \mid t), P(R_s \mid t)) - (1 - P(s, s_{ml} \mid t)))}{\min(P(L_s \mid t), P(R_s \mid t))}$$

where $1 - P(s, s_{ml})$ is the probability of error of the split s_{ml} when it is used to predict the split s. Thus, the predictive measure of association measures the relative difference between the probability of error of the split s' and the probability of error of the split s_{ml}. The bigger the difference, the higher the value of the predictive measure of association and the better the surrogate split s_{ml}. $A(s, s_{ml})$ can take negative values for some splits only if they do worse than the split s'. Thus, these splits clearly shouldn't be used as surrogate splits.

Among the $m - 1$ (or less if some were worse than s') surrogate splits, the split with the highest value of $A(s, s_{ml})$ is called the first surrogate split. The second surrogate split is the split with the second highest such value. Similarly, the third, fourth,... surrogate splits can be defined.

As stated previously, we can obtain up to $m - 1$ surrogate splits, one for each predictor variable other than the variable on which the split s is based. If the value of the variable on which the split s is based is missing for some observations, the split s can be replaced by the first surrogate split. If the value of the variable on which this split is based is also missing for some observations, the second surrogate split is then used and so on. If none of the surrogate splits can be used, the split s' is used. Thus, all the observations can be classified by a split at some point.

2.5 Misclassification costs

In some problems, it is possible that it is worse to misclassify some observations in some classes than in others. For example, suppose we have with a response variable containing 3 classes denoted class 1, class 2 and class 3. Then, it is possible that it is worse to misclassify class 1 observations in class 3 than in class 2. In such cases, the model can be improved by using misclassification costs.

Misclassification costs represent the cost of classifying an observation that is truly in some class j in another class, say class i. For example, in our problem, it is worse to misclassify a fatal collision as a non-fatal injury collision than to misclassify a non-fatal injury collision as a fatal collision, so the cost for the first case should be higher than the cost for the second case. The higher the costs, the more important it is to classify the corresponding observations correctly.

Misclassifications costs are represented by a matrix of the form:

$$C = \begin{bmatrix} 0 & c_{21} & \cdots & c_{k1} \\ c_{12} & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & c_{k,k-1} \\ c_{1k} & \cdots & c_{k-1,k} & 0 \end{bmatrix}$$

where c_{ij}, $i = 1, \ldots, J$, $j = 1, \ldots, J$, represents the cost of classifying a class j observation in class i. The cost of classifying an observation correctly is clearly 0. If the other costs are not specified, they are usually all taken to be 1. That is, all observations are equally important to classify correctly.

When misclassification costs are introduced in a model, it is no longer useful to find a tree minimizing the probability of misclassification. We would like to find a tree minimizing the total misclassification cost, also called the expected cost of misclassification (ECM), which takes into account the misclassification cost. The expected misclassification cost for a tree can be generalized from equation 2.2.1 by incorporating costs in the formula. In this case, it is given by

$$ECM = \sum_t \sum_j c_{ij} P_{Y|A}(j \mid t) \times P_A(t) \tag{2.5.1}$$

where i represents the most likely class for the node t and the summation in t is over the terminal nodes.

The expected misclassification cost can also be given by the equivalent equation

$$ECM = \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij} Q(i \mid j) P_Y(j)$$

where $Q(i \mid j)$ is the probability that an observation is classified in class i by the tree given that it is really in class j.

Remember that the Gini criterion gave an estimate of the node probability of misclassification, only with better properties. Thus, intuitively, we would try to estimate the expected cost of misclassification for a node by a generalized version of the Gini criterion taking into account the misclassification costs.

Since it was shown that the Gini criterion is given by

$$G_t = \sum_{k \neq j} P_{Y|A}(j \mid t) P_{Y|A}(k \mid t),$$

the Gini criterion is generalized by

$$G_t = \sum_{k,j} c_{kj} P_{Y|A}(j \mid t) P_{Y|A}(k \mid t).$$

The Gini criterion satisfied the four conditions to be an impurity criterion. However, this generalized version of the Gini criterion may not always be a concave function. Thus, it can't be used as an impurity criterion.

A way to resolve this problem is to alter or modify the prior probabilities in such a way that we will be able to use the regular Gini criterion or deviance as an impurity criterion.

Let C' be a multiple of the unit misclassification cost matrix. A multiple of the unit cost matrix is equivalent to the unit cost matrix itself. We can explain this as follows. The goal of using misclassification cost matrices is to take into account the differences in the importance of classifying correctly observations from some classes compared to observations in other classes. However, using a multiple of the unit misclassification cost matrix doesn't change anything in the sense that the misclassification costs remain equal for all observations. Thus, the choices of splits are the same using both matrices.

C' is given by

$$C' = a \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \vdots & \vdots \\ \vdots & \vdots & 0 & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}$$

To be able to use the regular Gini criterion we need to alter the prior probabilities so that even though the unit misclassification cost matrix is used instead of C, the expected cost of misclassification remains the same.

Remember that the expected cost of misclassification is given by

$$ECM = \sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij} Q(i \mid j) P_Y(j).$$

Thus, for the expected misclassification cost to remain the same, the modified prior probabilities $P'_Y(1), \dots, P'_Y(J)$ must satisfy

$$\sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij} Q(i \mid j) P_Y(j) = \sum_{i=1}^{J} \sum_{j=1}^{J} a \times c'_{ij} Q(i \mid j) P'_Y(j).$$

This is equivalent to saying that the altered prior probabilities must satisfy

$$c_{ij} Q(i \mid j) P_Y(j) = a \times c'_{ij} Q(i \mid j) P'_Y(j), \forall\ i, j,\ i \neq j.$$

However, since $Q(i \mid j)$ remains the same, the altered prior probabilities must satisfy only

$$c_{ij} P_Y(j) = a \times c'_{ij} P'_Y(j), \forall\ i, j,\ i \neq j.$$

Let's first look at the situation where the original costs are such that $c_{ij} = c_j$, that is, the cost of misclassifying an observation from class j in any other class i is the same.

In such a case, the altered prior probabilities must satisfy,

$$c_j P_Y(j) = a \times c'_{ij} P'_Y(j), \forall\ i, j,\ i \neq j.$$

However, as C' is the unit cost matrix $c'_{ij} = 1, \forall\ i \neq j$. Thus, $\forall\ j$ we have:

$$P'_Y(j) = \frac{c_j}{a} \times P_Y(j).$$

It remains to find what the value of a should be. As $P'_Y(1), \ldots, P'_Y(J)$ are probabilities, they should be between 0 and 1. Moreover, they should add up to 1. Thus, a should be taken to be the sum of all the terms of the form $c_j P_Y(j), j = 1, \ldots, J$, as this would satisfy these two conditions.

So, taking the prior probabilities to be

$$P'_Y(j) = \frac{c_j P_Y(j)}{\sum_{j=1}^{J} c_j P_Y(j)}$$

and using the misclassification cost matrix C' produces the same expected cost of misclassification than when using the original prior probabilities and misclassification costs for any given data set. However, we now have a problem with no misclassification costs.

Now, let's look at the general case where $c_{ij} \neq c_j$ for some or all j. In this case, the cost of misclassifying a class j observation is not the same for all classes i in which the observation can be misclassified. However, in such a case, it is impossible to obtain an exact vector of altered prior probabilities. Nonetheless, we would like to find prior probabilities $P'_Y(1), \ldots, P'_Y(J)$ such that the expected cost of misclassification calculated using the altered prior probabilities is as close as possible to the true expected cost of misclassification. To achieve this, we estimate c_j, the cost of misclassifying an observation from class j, by its mean over the $J-1$ classes in which the observation can be misclassified. That is,

$$c_j = \frac{1}{J-1} \sum_{i=1}^{J-1} c_{ij}.$$

Then, the altered prior probabilities are calculated as before.

Thus, in both cases, the problem is now a problem with no misclassification costs, but with different priors. Hence, it can be handled as all other problems where no misclassification costs are used. That is, the usual impurity criteria, the regular Gini criterion or the deviance can be used to obtain the splits. The impurity is thus calculated using these new priors and unit misclassification costs.

However, note that when calculating the exact expected cost of misclassification of a tree for a given data set, using equation (2.5.1), the original misclassification costs and prior probabilities are used.

Note also that when we prune a tree which uses misclassification costs, the cost of the tree becomes the expected cost of misclassification instead of the probability of misclassification.

2.6 Example on a small data set

An example on a small data set will help to understand the mechanics of classification trees. A data set containing 40 observations for which the value of the response variable and of 3 predictor variables are given was created.

The response variable Y contains 2 categories: 1 and 2. Category 1 contains 25 observations and category 2 contains 15 observations. Predictor variable X_1 is a numerical variable taking values between 2 and 99 in this particular data set, but it can take any numerical value in general. Variable X_2 is an ordered categorical variable divided into 3 categories: 1, 2 and 3, containing respectively 13, 14 and 13 observations. Variable X_3 is a categorical variable containing 2 categories: A and B, containing respectively 19 and 17 observations. Note that X_3 has 4 missing values.

The prior class probabilities are taken to be the empirical prior probabilities of each class based on the sample. That is, $P_Y(1) = 0.625$ and $P_Y(2) = 0.375$ (we assume that these empirical probabilities are representative of the population proportions for each class, so that they can be used as prior probabilities). The Gini criterion will be used as the impurity criterion. However, the same steps are used when growing a tree with deviance as the impurity criterion, except that it can produce slightly different splits at some points.

A cost of $c_{12} = 8$ is put on misclassifying a class 2 observation in class 1 and a cost of $c_{21} = 5$ is put on misclassifying a class 1 observation in class 2. That is, misclassifying a class 2 observation is 1.6 times worse than misclassifying a class 1

observation.

A tree is grown from the given data. The root node of the tree, called node 1, contains the 40 observations in the data set. If we were to assign the observations to a class at this point, we would assign them to the class such that the expected cost of misclassification produced is the lowest. In this particular case, as there is only two categories to our predictor variable, if we classified the observations in category 1, the ECM would be the cost of misclassifying an observation from class 2 in class 1. Thus, ECM would be

$$ECM_1 = c_{12}P_{Y|A}(2 \mid 1) = 8 \times \frac{15}{40} = 3$$

Remember that if the prior probabilities are the empirical ones, $P_{Y|A}(j \mid t) = n_{jt}/n_t$. Here, $P_{Y|A}(2 \mid 1) = n_{21}/n_1$ where $n_{21} = 15$ and $n_1 = 40$.

Similarly, if we classified the observations in category 2, the ECM would be

$$ECM_1 = c_{21}P_{Y|A}(1 \mid 1) = 5 \times \frac{25}{40} = 3.125$$

as $n_{11} = 25$ and $n_1 = 40$.

As the expected cost of misclassification is lower when classifying the observations in class 1, all the observations would be classified in class 1 if the root node only was used to classify the observations. The goal of growing a tree is to improve this ECM.

Trees are grown by splitting nodes based on the values of the predictor variables. We first split node 1. The observations in this node can be split based on one of the three predictor variables: X_1, X_2 or X_3.

As variable X_1 is a numerical variable, many splits based on X_1 are possible. However, as splits should be binary in classification trees, they should be generated by questions of the form "Is $X_1 < a$?" where $a \in R$. For each observation, if $X_1 < a$, such a split sends the observation to the left child of node 1, called node 2. Otherwise, the split sends the observation to the right child of node 1, node 3. Note that some software use splits of this form and their complementary splits, generated by questions of the form "Is $X_1 > a$?", in an interchangeable manner. The use of these

complementary splits produce the same results, except that left and right child nodes are inverted.

Any value $a \in R$ can be used as a split point. However, remember that the possible split points for numerical variables are the midpoints between each pair of consecutive values present in the data set. In the current example, as 47 and 48 are two consecutive values present in the data set, the question "Is $X_1 < 47.5$?" would be used to generate the split sending the observations for which $X_1 \leq 47$ to the left child node and those for which $X_1 \geq 48$ to the right child node. So, there is a finite number of splits for X_1.

As X_2 is an ordered categorical variable, only two splits are possible based on this variable as the ordering of the categories need to be taken into account. These two splits are generated from the questions: "Is $X_2 \in \{1\}$?" and "Is $X_2 \in \{1,2\}$?", respectively. A split such as the one given by the question "Is $X_2 \in \{1,3\}$?" would not be possible as it groups categories which are not consecutive in the determined ordering of the categories, in this case 1,2 and 3.

X_3 is a categorical variable with only two categories, so only one split is possible based on this variable. This split is generated by the question "Is $X_1 \in \{A\}$?".

For each possible split, we need to calculate the impurity criterion. Remember that the Gini criterion is the impurity criterion used in this example. However, misclassification costs are used in this problem, and thus the prior class probabilities should be altered to take them into account in the calculation of the impurity using the Gini criterion. Remember that these costs are $c_{12} = 8$ and $c_{21} = 5$. The prior class probabilities are $P_Y(1) = 0.625$ and $P_Y(2) = 0.375$. As the cost of misclassifying the observations of a given class is the same for all observations (as there is only one class in which the observations can be misclassified), we find the exact altered prior class probabilities. The altered prior probabilities are given by

$$P'_Y(1) = \frac{c_1 P_Y(1)}{\sum_{j=1}^{2} c_j P_Y(j)} = \frac{5 \times 0.625}{5 \times 0.625 + 8 \times 0.375} = 0.51 \text{ and}$$

$$P'_Y(2) = \frac{c_2 P_Y(2)}{\sum_{j=1}^{2} c_j P_Y(j)} = \frac{8 \times 0.375}{5 \times 0.625 + 8 \times 0.375} = 0.49.$$

The Gini criterion is now calculated for both child nodes produced by each possible split, from which the total impurity of the children of node 1 can be calculated for each possible split. In a two-class problem, the Gini criterion for a node t is given by

$$G_t = P_{Y|A}(1 \mid t)P_{Y|A}(2 \mid t) + P_{Y|A}(2 \mid t)P_{Y|A}(1 \mid t) = 2P_{Y|A}(1 \mid t)P_{Y|A}(2 \mid t).$$

Moreover, remember that $P_{Y|A}(j \mid t) = \frac{\frac{n_{jt}}{n_j} \times P'_Y(j)}{\sum_j \frac{n_{jt}}{n_j} \times P'_Y(j)}$.

We'll look at a few examples of the calculation of the Gini criterion. We first look at the split generated by the question "Is $X_1 < 47.5$?". This split sends 21 observations to node 2, the left child of node 1 and 19 observations to node 3, the right child of node 1. Among the 21 observations of node 1 sent to node 2, 15 are in class 1 and 6 are in class 2 whereas among the 19 observations of node 1 sent to node 3, 10 are in class 1 and 9 are in class 2.

Thus, the value of the Gini criterion for nodes 2 and 3 are respectively

$$G_2 = 2 \times \frac{\frac{15}{25} \times 0.51}{\frac{15}{25} \times 0.51 + \frac{6}{15} \times 0.49} \times \frac{\frac{6}{15} \times 0.49}{\frac{15}{25} \times 0.51 + \frac{6}{15} \times 0.49} = 0.4758 \text{ and}$$

$$G_3 = 2 \times \frac{\frac{10}{25} \times 0.51}{\frac{10}{25} \times 0.51 + \frac{9}{15} \times 0.49} \times \frac{\frac{9}{15} \times 0.49}{\frac{10}{25} \times 0.51 + \frac{9}{15} \times 0.49} = 0.4838.$$

The total impurity of the children of node 1 is the sum of the two node impurities weighted by the proportion of observations from node 1 sent to each node, that is:

$$G_2 \times \frac{21}{40} + G_3 \times \frac{19}{40} = \left(0.4758 \times \frac{21}{40}\right) + \left(0.4838 \times \frac{19}{40}\right) = 0.4796.$$

The impurity of the tree consisting of node 1 only was

$$G_1 = 2 \times \frac{\frac{25}{25} \times 0.51}{\frac{25}{25} \times 0.51 + \frac{15}{15} \times 0.49} \times \frac{\frac{15}{15} \times 0.49}{\frac{25}{25} \times 0.51 + \frac{15}{15} \times 0.49} = 0.4998.$$

Thus, this split leads to a decrease in the total tree impurity. However, we need to find the value of the Gini criterion for all other possible splits of node 1 in order to

find which one leads to the biggest decrease in the tree impurity. The Gini criterion is calculated similarly for all other splits based on X_1 and those based on X_2.

However, the impurity is calculated somewhat differently for the splits based on variable X_3 as some observations have missing X_3 values. Only the observations for which the value of X_3 is available are used to calculate the impurity associated with the splits based on X_3.

Two splits based on the variable X_3 exists, the split generated by the question "Is $X_3 \in \{A\}$?" and its complement, generated by the question "Is $X_3 \in \{B\}$?". The value of X_3 is missing for 4 of the 40 observations in this data set. Thus, the calculation of the impurity of the nodes produced by these splits is based only on 36 observations. The split generated by the question "Is $X_3 \in \{A\}$?" sends 19 observations to node 2 and 17 observations to node 3. Among the 19 observations of node 1 sent to node 2 by this split, 10 are in class 1 and 9 are in class 2 whereas among the 17 observations of node 1 sent to node 3, 14 are in class 1 and 3 are in class 2.

Thus, for this split, the values of the Gini criterion for nodes 2 and 3 are respectively

$$G_3 = 2 \times \frac{\frac{10}{24} \times 0.51}{\frac{10}{24} \times 0.51 + \frac{9}{12} \times 0.49} \times \frac{\frac{9}{12} \times 0.49}{\frac{10}{24} \times 0.51 + \frac{9}{12} \times 0.49} = 0.4591 \text{ and}$$

$$G_3 = 2 \times \frac{\frac{14}{24} \times 0.51}{\frac{14}{24} \times 0.51 + \frac{3}{12} \times 0.49} \times \frac{\frac{3}{12} \times 0.49}{\frac{14}{24} \times 0.51 + \frac{3}{12} \times 0.49} = 0.4872.$$

and thus, the total impurity of the children of node 1 is

$$G_2 \times \frac{19}{36} + G_3 \times \frac{17}{36} = \left(0.4591 \times \frac{19}{36}\right) + \left(0.4872 \times \frac{17}{36}\right) = 0.4723.$$

However, the split of node 1 producing the greatest impurity reduction is the split generated by the question "Is $X_1 < 52$?". We will denote this split by s_1. s_1 sends 24 observations to node 2 and 16 observations to node 3. Among the 24 observations from node 1 sent to node 2 by s_1, 18 are in class 1 and 6 are in class 2 whereas among the 16 observations of node 1 sent to node 3, 7 are in class 1 and 9 are in class 2.

Thus, for the split s_1, the values of the Gini criterion for nodes 2 and 3 are respectively

$$G_2 = 2 \times \frac{\frac{18}{25} \times 0.51}{\frac{18}{25} \times 0.51 + \frac{6}{15} \times 0.49} \times \frac{\frac{6}{15} \times 0.49}{\frac{18}{25} \times 0.51 + \frac{6}{15} \times 0.49} = 0.4538 \text{ and}$$

$$G_3 = 2 \times \frac{\frac{7}{25} \times 0.51}{\frac{7}{25} \times 0.51 + \frac{9}{15} \times 0.49} \times \frac{\frac{9}{15} \times 0.49}{\frac{7}{25} \times 0.51 + \frac{9}{15} \times 0.49} = 0.4401.$$

The total impurity of the children of node 1 is thus

$$G_2 \times \frac{24}{40} + G_3 \times \frac{16}{40} = \left(0.4538 \times \frac{24}{40}\right) + \left(0.4401 \times \frac{16}{40}\right) = 0.4483,$$

the smallest impurity produced by any split of node 1.

To be able to classify observations with missing X_1 values in future data sets, we must find the surrogate splits for this split, if such splits exists. Surrogate splits are splits based on the other predictor variables. The surrogate splits for s_1 are obtained by first finding, for each of the variables X_2 and X_3, the splits with the highest proportion of observations sent to the same node as by s_1. Then, these splits are ordered based on their measure of predictive association. The split with the highest measure is the first surrogate split and so on.

Four splits based on X_2 are available. These splits are generated by questions "Is $X_2 \in \{1\}$?", "Is $X_2 \in \{1,2\}$?", "Is $X_2 \in \{2,3\}$?" and "Is $X_2 \in \{3\}$?". Note that the last two splits are the complements of the first two. We denote these four splits by $s_{121}, s_{122}, s_{123}$ and s_{124}.

As 24 observations are sent to node 2 by s_1 and 7 of these observations are also sent to the node 2 by s_{121} and 16 observations are sent to node 3 by s_1 and 10 of these are also sent to the node 3 by s_{121}, the probability that s_1 sends an observation in node 1 to the same node as s_{121} is

$$P(s_1, s_{121} \mid 1) = \frac{7+10}{40} = 0.425.$$

Similarly, the probabilities that s_{122}, s_{123} and s_{124} send an observation to the same child node as s_1 are respectively

$$P(s_1, s_{122} \mid 1) = \frac{15+5}{40} = 0.5$$

$$P(s_1, s_{123} \mid 1) = \frac{17+6}{40} = 0.575.$$
$$P(s_1, s_{124} \mid 1) = \frac{9+11}{40} = 0.5.$$

The same probability is also calculated for the two splits based on the variable X_3, the split generated by the question "Is $X_3 \in \{A\}$?" and its complementary split, generated by "Is $X_3 \in \{B\}$?" (based on the observations for which X_3 is known). We denote these splits s_{131} and s_{132}. We choose the splits, based on X_2 and X_3 respectively, with the highest probabilities. The splits s_{123} and s_{131} are chosen. For the split s_{131}, the probability is

$$P(s_1, s_{131} \mid 1) = \frac{14+8}{36} = 0.611.$$

We now find the measure of predictive association for each of these two splits. Remember that s_1 was sending 24 observations to node 2 and 16 to node 3. So, $P(L_{s_1} \mid 1) = 0.6$ and $P(R_{s_1} \mid 1) = 0.4$. Thus,

$$\begin{aligned} A(s_1, s_{123}) &= \frac{\min(P(L_{s_1} \mid 1), P(R_{s_1} \mid 1)) - (1 - P(s_1, s_{123}))}{\min(P(L_{s_1} \mid 1), P(R_{s_1} \mid 1))} \\ &= \frac{0.4 - 0.425}{0.4} = -0.025 \end{aligned}$$

$$\begin{aligned} A(s_1, s_{131}) &= \frac{\min(P(L_{s_1} \mid 1), P(R_{s_1} \mid 1)) - (1 - P(s_1, s_{131}))}{\min(P(L_{s_1} \mid 1), P(R_{s_1} \mid 1))} \\ &= \frac{0.4 - 0.389}{0.4} = 0.03. \end{aligned}$$

The measure of predictive association is close to 0 or smaller than 0 in both cases. Thus, none of these is a good surrogate split, and thus the split s', which sends all the observations to node 2, as $P(L_{s_1} \mid 1) > P(R_{s_1} \mid 1)$, is used when the value of X_1 is missing.

The tree now contains 3 nodes. As the two terminal nodes are not pure and as they contain more than 10 observations each (the number of observations chosen as the threshold for the minimum number of observations needed in a node to continue splitting), we continue to split both nodes.

Node 2 contains 24 observations (18 from class 1 and 6 from class 2). We split node 2 based on these observations. As we did for node 1, the total impurity of the child nodes for each possible split of node 2 is obtained. Note that there are fewer splits to check as $X_1 < 52$ for all the observations in node 2. Thus, all splits having split points at greater values of X_1 are not useful anymore.

The split of node 2 producing the greatest impurity reduction is the split given by the rule "Is $X_3 \in \{B\}$?". We denote this split by s_2. s_2 sends 9 observations to node 4, the left child of node 2, and 14 observations to node 5, the right child of node 2, whereas 1 observation has a missing X_3 value. Thus, only 23 observations will be used to calculate the impurity of the children of node 2. The 9 observations of node 2 sent to node 4 are in class 1 whereas among the 14 observations from node 2 sent to node 5, 8 are in class 1 and 6 are in class 2.

Thus, the values of the Gini criterion for nodes 4 and 5 are respectively

$$G_4 = 2 \times \frac{\frac{9}{25} \times 0.51}{\frac{9}{25} \times 0.51 + \frac{0}{15} \times 0.49} \times \frac{\frac{0}{15} \times 0.49}{\frac{9}{25} \times 0.51 + \frac{0}{15} \times 0.49} = 0 \text{ and}$$

$$G_5 = 2 \times \frac{\frac{8}{25} \times 0.51}{\frac{8}{25} \times 0.51 + \frac{6}{15} \times 0.49} \times \frac{\frac{6}{15} \times 0.49}{\frac{8}{25} \times 0.51 + \frac{6}{15} \times 0.49} = 0.4958.$$

Node 4 is a pure node as all its observations are in class 1, so we won't need to split it further. The total impurity of the children of node 2 is

$$G_4 \times \frac{9}{23} + G_5 \times \frac{14}{23} = \left(0 \times \frac{9}{23}\right) + \left(0.4958 \times \frac{14}{23}\right) = 0.3018.$$

We now find the surrogate splits for s_2, based on the variables X_1 and X_2 in this case.

Many splits based on X_1 are available. We denote these splits by $s_{211}, s_{212}, \ldots$. We calculate the probability that each of these splits sends an observation to the same child of node 2 as the split s_2, based on the observations in node 2.

The split based on X_1 with the highest probability is the split s_{2163}, generated by the question "Is $X_1 > 46$?". As 9 observations are sent to node 4 by s_2 and 3 of these are also sent to node 4 by s_{2163} and 14 observations are sent to the node 5 by

s_2 and 13 of these are also sent to node 5 by s_{2163}, the probability that s_2 sends an observation to the same node as s_{2163} is

$$P(s_2, s_{2163} \mid 2) = \frac{3+13}{23} = 0.696.$$

The same probability is also calculated for the splits based on X_2 (the same splits as for node 1). We denote these splits by $s_{221}, s_{222}, s_{223}$ and s_{224}. The split based on X_2 with the highest probability is the split s_{221}, generated by the question "Is $X_2 \in \{1\}$?". As 9 observations are sent to node 4 by s_2 and 4 of these are also sent to node 4 by s_{221} and 14 observations are sent to the node 5 by s_2 and 10 of these are also sent to node 5 by s_{221}, the probability that s_2 sends an observation to the same node as s_{221} is

$$P(s_2, s_{221} \mid 2) = \frac{4+10}{23} = 0.609.$$

We now find the measure of predictive association for both splits. Remember that the split s_2 was sending 9 observations to node 4 and 14 to node 5, so that $P(L_{s_2} \mid 2) = 0.391$ and $P(R_{s_2} \mid 2) = 0.609$. Thus,

$$\begin{aligned} A(s_2, s_{2163}) &= \frac{\min(P(L_{s_2} \mid 2), P(R_{s_2} \mid 2)) - (1 - P(s_2, s_{2163}))}{\min(P(L_{s_2} \mid 2), P(R_{s_2} \mid 2))} \\ &= \frac{0.391 - 0.304}{0.391} = 0.222 \end{aligned}$$

$$\begin{aligned} A(s_2, s_{221}) &= \frac{\min(P(L_{s_2} \mid 2), P(R_{s_2} \mid 2)) - (1 - P(s_2, s_{221}))}{\min(P(L_{s_2} \mid 2), P(R_{s_2} \mid 2))} \\ &= \frac{0.391 - 0.391}{0.391} = 0. \end{aligned}$$

As the measure of predictive association is 0 for s_{221}, we don't use it as a surrogate split. However, if the value of X_1 is available for an observation, we use the split s_{2163} to choose where to send the observation. Otherwise s' is used. Here s' sends all the observations to node 5 as $P(R_{s_2} \mid 2) > P(L_{s_2} \mid 2)$. In particular, we can use our surrogate split to determine where to send the observation in node 2 for which X_3 is not available. As $X_1 < 46$ for this observation, we send it to node 5.

Node 3 contains 16 observations (7 from class 1 and 9 from class 2). We split node 3 based on these observations. As before, the total impurity of the child nodes produced by each possible split of node 3 is obtained. Note that there are fewer splits to check as $X_1 \geq 52$ for all observations in node 3.

The split of node 3 producing the greatest impurity reduction is the split generated by the question "Is $X_2 \in \{1\}$?". We will denote this split by s_3. s_3 sends 6 observations to node 6, the left child of node 3, and 10 observations to node 7, the right child of node 3. Among the 6 observations from node 3 sent to node 6 by s_3, 4 are in class 1 and 2 are in class 2 whereas among the 10 observations of node 3 sent to node 7, 3 are in class 1 and 7 are in class 2.

Thus, for this split, the values of the Gini criterion for nodes 6 and 7 are respectively

$$G_6 = 2 \times \frac{\frac{4}{25} \times 0.51}{\frac{4}{25} \times 0.51 + \frac{2}{15} \times 0.49} \times \frac{\frac{2}{15} \times 0.49}{\frac{4}{25} \times 0.51 + \frac{2}{15} \times 0.49} = 0.4940 \text{ and}$$

$$G_7 = 2 \times \frac{\frac{3}{25} \times 0.51}{\frac{3}{25} \times 0.51 + \frac{7}{15} \times 0.49} \times \frac{\frac{7}{15} \times 0.49}{\frac{3}{25} \times 0.51 + \frac{7}{15} \times 0.49} = 0.3330.$$

As node 6 and 7 contain 6 and 10 observations respectively, we won't need to split them further. The total impurity of the children of node 3 is

$$G_6 \times \frac{6}{16} + G_7 \times \frac{10}{16} = \left(0.4940 \times \frac{6}{16}\right) + \left(0.3330 \times \frac{10}{16}\right) = 0.3934.$$

We now find the surrogate splits for s_3. In this case, they will be based on X_1 and X_3.

The splits based on X_1 are denoted respectively by $s_{311}, s_{312}, \ldots$. We calculate the probability that each of these splits sends an observation to the same child of node 3 as the split s_3, based on the observations in node 3.

The split based on X_1 with the highest probability is s_{3175}, generated by the question "Is $X_1 > 91$?". As 6 observations are sent to node 6 by s_3 and 2 of these are also sent to node 6 by s_{3175} and 10 observations are sent to node 7 both by s_3

and s_{3175}, the probability that s_3 sends an observation in node 3 to the same node as s_{3175} is

$$P(s_3, s_{3175} \mid 3) = \frac{2+10}{16} = 0.75.$$

The same probability is also calculated for both splits based on X_3. The split based on X_3 with the highest probability is s_{332}, generated by the question "Is $X_2 \in \{1,2\}$?". As 6 observations are sent to node 6 by s_3 and 4 of these are also sent to node 6 by s_{332} and 10 observations are sent to node 7 by s_3 and 4 of these are also sent to node 7 by s_{332}, the probability that s_3 sends an observation to the same node as s_{332} is

$$P(s_3, s_{332} \mid 3) = \frac{4+4}{13} = 0.615.$$

We now find the measure of predictive association for both splits. As s_3 was sending 6 observations to node 6 and 10 observations to node 7, so that $P(L_{s_3} \mid 3) = 0.375$ and $P(R_{s_3} \mid 3) = 0.625$, the measures of predictive association for these splits are respectively

$$\begin{aligned} A(s_3, s_{3175}) &= \frac{\min(P(L_{s_3} \mid 3), P(R_{s_3} \mid 3)) - (1 - P(s_3, s_{3175}))}{\min(P(L_{s_3} \mid 3), P(R_{s_3} \mid 3))} \\ &= \frac{0.375 - 0.25}{0.375} = 0.333 \end{aligned}$$

$$\begin{aligned} A(s_3, s_{332}) &= \frac{\min(P(L_{s_3} \mid 3), P(R_{s_3} \mid 3)) - (1 - P(s_3, s_{332}))}{\min(P(L_{s_3} \mid 3), P(R_{s_3} \mid 3))} \\ &= \frac{0.375 - 0.385}{0.375} = -0.027. \end{aligned}$$

As the measure of predictive association is smaller than 0 for s_{332}, we don't use it as a surrogate split. Thus, if the value of X_1 is available for an observation, we use the split s_{3175} to determine where to send it when X_2 is missing. Otherwise s' is used. Here, s' sends all the observations to node 7 as $P(R_{s_3} \mid 3) > P(L_{s_3} \mid 3)$.

Among the 4 terminal nodes we have at this point, the only node we continue to split is node 5. Node 5 contains 15 observations (9 from class 1 and 6 from class 2). We split it based on these observations. We first obtain the total impurity of the

child nodes produced by each possible split of node 5. Note that for all observations in node 5, $X_1 < 52$ and $X_3 = A$. Thus, fewer splits are available. In particular, we cannot split on the variable X_3

The split of node 5 producing the greatest impurity reduction is the split generated by the question "Is $X_1 < 17$?". We denote this split by s_4. s_4 sends 3 observations to node 10, the left child of node 5, and 12 observations to node 11, the right child of node 5. The 3 observations from node 5 sent to node 10 by s_4 are in class 1 whereas among the 12 observations of node 5 sent to node 11, 6 are in class 1 and 6 are in class 2.

Thus, for this split, the values of the Gini criterion for nodes 10 and 11 are respectively

$$G_{10} = 2 \times \frac{\frac{3}{25} \times 0.51}{\frac{3}{25} \times 0.51 + \frac{0}{15} \times 0.49} \times \frac{\frac{3}{15} \times 0.49}{\frac{3}{25} \times 0.51 + \frac{0}{15} \times 0.49} = 0 \text{ and}$$

$$G_{11} = 2 \times \frac{\frac{6}{25} \times 0.51}{\frac{6}{25} \times 0.51 + \frac{6}{15} \times 0.49} \times \frac{\frac{6}{15} \times 0.49}{\frac{6}{25} \times 0.51 + \frac{6}{15} \times 0.49} = 0.4731.$$

Node 10 is a pure node, so we don't split it further. However, node 11 is not pure and contains 12 observations, so we continue to split it. So, the total impurity of the children of node 5 is

$$G_{10} \times \frac{3}{15} + G_{11} \times \frac{12}{15} = \left(0 \times \frac{3}{15}\right) + \left(0.4731 \times \frac{12}{15}\right) = 0.3785.$$

We now find the surrogate split for s_4. Only one surrogate split, a split based on X_2, will be obtained,as s_4 is based on X_1 and all observations have the same X_3 value.

Four splits based on X_2 are available, as usual. We denote these splits s_{421}, s_{422}, s_{423} and s_{424}. We calculate the probability that each of these splits sends an observation to the same child node of node 5 as s_4, based on the observations in node 5.

The split with the highest probability is the split s_{424}, generated by the question "Is $X_2 \in \{2, 3\}$?". As 3 observations are sent to node 10 by s_4 and 2 of these are also sent to node 10 by s_{424} and 12 observations are sent to node 11 by s_4 and 10 of these

are also sent to node 11 by s_{424}, the probability that s_4 sends an observation to the same node as s_{424} is

$$P(s_4, s_{424} \mid 5) = \frac{2+10}{15} = 0.8.$$

As s_4 was sending 3 observations to node 10 and 12 to node 11, so that $P(L_{s_4} \mid 5) = 0.2$ and $P(R_{s_4} \mid 5) = 0.8$, the measure of predictive association of this split is

$$A(s_4, s_{424}) = \frac{\min(P(L_{s_4} \mid 5), P(R_{s_4} \mid 5)) - (1 - P(s_4, s_{424}))}{\min(P(L_{s_4} \mid 5), P(R_{s_4} \mid 5))} = \frac{0.2 - 0.2}{0.2} = 0.$$

As the measure of predictive association is 0 for s_{424}, we don't use it as a surrogate split. Instead, we use s'. In this case, as $P(R_{s_4} \mid 5) > P(L_{s_4} \mid 5)$, s' sends the observations with unknown X_1 value to node 11.

At this point, the only node we continue to split is node 11. Node 11 contains 12 observations (6 from class 1 and 6 from class 2). We split node 11 based on these observations. Again, we find the total impurity of the child nodes produced by each possible split. Note that among the observations in node 11, $17 < X_1 < 52$ and $X_3 = A$, so fewer splits are available. In particular, we cannot split on the variable X_3.

The split of node 11 producing the greatest impurity reduction is the split generated by the question "Is $X_1 > 41$?". We denote this split by s_5. s_5 sends 5 observations to node 22, the left child of node 11, and 7 observations to node 23, the right child of node 11. Among the 5 observations from node 11 sent to node 22 by s_5, 4 are in class 1 and 1 is in class 2, whereas among the 7 observations of node 11 sent to node 23, 2 are in class 1 and 5 are in class 2.

Thus, the values of the Gini criterion for nodes 22 and 23 are respectively

$$G_{22} = 2 \times \frac{\frac{4}{25} \times 0.51}{\frac{4}{25} \times 0.51 + \frac{1}{15} \times 0.49} \times \frac{\frac{1}{15} \times 0.49}{\frac{4}{25} \times 0.51 + \frac{1}{15} \times 0.49} = 0.4084 \text{ and}$$

$$G_{23} = 2 \times \frac{\frac{2}{25} \times 0.51}{\frac{2}{25} \times 0.51 + \frac{5}{15} \times 0.49} \times \frac{\frac{5}{15} \times 0.49}{\frac{2}{25} \times 0.51 + \frac{5}{15} \times 0.49} = 0.32.$$

Thus, the total impurity of the children of node 11 is

$$G_{22} \times \frac{5}{12} + G_{23} \times \frac{7}{12} = \left(0.4084 \times \frac{5}{12}\right) + \left(0.32 \times \frac{7}{12}\right) = 0.3569.$$

We now find the surrogate split for s_5. As s_5 is based on X_1 and all observations in node 11 satisfy $X_3 = A$, the only surrogate split will be based on X_2.

Again, the same splits based on X_2 are available. We denote these splits by $s_{521}, s_{522}, s_{523}$ and s_{524}. We calculate the probability that each of these splits sends an observation to the same child of node 11 as s_5, based on the observations in node 11.

The split with the highest probability is s_{521}, generated by the question "Is $X_2 \in \{1\}$?". As 5 observations are sent to node 22 by s_5 and 2 of these are also sent to node 22 by s_{521} and 7 observations are sent to node 23 by s_5 and 6 of these are also sent to node 23 by s_{521}, the probability that s_5 sends an observation to the same node as s_{521} is

$$P(s_5, s_{521} \mid 11) = \frac{2+6}{12} = 0.667.$$

As s_5 was sending 5 observations to node 22 and 7 to node 23, so that $P(L_{s_5} \mid 11) = 0.42$ and $P(R_{s_5} \mid 11) = 0.58$, the measure of predictive association of this split is

$$\begin{aligned} A(s_5, s_{521}) &= \frac{\min(P(L_{s_5} \mid 11), P(R_{s_5} \mid 11)) - (1 - P(s_5, s_{521}))}{\min(P(L_{s_5} \mid 11), P(R_{s_5} \mid 11))} \\ &= \frac{0.42 - 0.333}{0.42} = 0.2. \end{aligned}$$

As the measure of predictive association is greater than 0, the split s_{521} can be used as a surrogate split for observations where X_1 is unknown but X_2 is known. Otherwise, we use s'. In this case, as $P(R_{s_5} \mid 11) > P(L_{s_5} \mid 11)$, s' sends the observations with unknown X_1 value to node 23.

As each terminal node now contains less than 10 observations or is pure, we don't need to split any node further.

The next step is to assign the observations to classes. For each terminal node, we assign the observations to the class such that the ECM is the lower. The terminal nodes are nodes 4, 6, 7, 10, 22 and 23.

For node 4, if we classify the observations in class 1, the ECM is

$$ECM_4 = c_{12} P_{Y|A}(2 \mid 4) = 8 \times \frac{0}{9} = 0$$

as $n_{24} = 0$ and $n_4 = 9$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_4 = c_{21} P_{Y|A}(1 \mid 4) = 5 \times \frac{9}{9} = 5$$

as $n_{14} = 9$ and $n_4 = 9$.

As the expected cost of misclassification is lower when classifying the observations in class 1, all the observations in node 4 are classified as being in class 1.

For node 6, if we classify the observations in class 1, the ECM is

$$ECM_6 = c_{12} P_{Y|A}(2 \mid 6) = 8 \times \frac{2}{6} = 2.667$$

as $n_{26} = 2$ and $n_6 = 6$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_6 = c_{21} P_{Y|A}(1 \mid 6) = 5 \times \frac{4}{6} = 3.333$$

as $n_{16} = 4$ and $n_6 = 6$.

As the expected cost of misclassification is lower when classifying the observations in class 1, all the observations in node 6 are classified as being in class 1.

For node 7, if we classify the observations in class 1, the ECM is

$$ECM_7 = c_{12} P_{Y|A}(2 \mid 7) = 8 \times \frac{7}{10} = 5.6$$

as $n_{27} = 7$ and $n_7 = 10$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_7 = c_{21} P_{Y|A}(1 \mid 7) = 5 \times \frac{3}{10} = 1.5$$

as $n_{17} = 3$ and $n_7 = 10$.

As the expected cost of misclassification is lower when classifying the observations in class 2, all the observations in node 7 are classified as being in class 2.

For node 10, if we classify the observations in class 1, the ECM is

$$ECM_{10} = c_{12} P_{Y|A}(2 \mid 10) = 8 \times \frac{0}{3} = 0$$

as $n_{210} = 0$ and $n_{10} = 3$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_{10} = c_{21} P_{Y|A}(1 \mid 10) = 5 \times \frac{3}{3} = 5$$

as $n_{110} = 3$ and $n_{10} = 3$.

As the expected cost of misclassification is lower when classifying the observations in class 1, all the observations in node 10 are classified as being in class 1.

For node 22, if we classify the observations in class 1, the ECM is

$$ECM_{22} = c_{12} P_{Y|A}(2 \mid 22) = 8 \times \frac{1}{5} = 1.6$$

as $n_{222} = 1$ and $n_{22} = 5$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_{22} = c_{21} P_{Y|A}(1 \mid 22) = 5 \times \frac{4}{5} = 4$$

as $n_{122} = 4$ and $n_{22} = 5$.

As the expected cost of misclassification is lower when classifying the observations in class 1, all the observations in node 22 are classified as being in class 1.

Finally, for node 23, if we classify the observations in class 1, the ECM is

$$ECM_{23} = c_{12} P_{Y|A}(2 \mid 23) = 8 \times \frac{5}{7} = 5.712$$

as $n_{223} = 5$ and $n_{23} = 7$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_{23} = c_{21} P_{Y|A}(1 \mid 23) = 5 \times \frac{2}{7} = 1.42$$

as $n_{123} = 2$ and $n_{23} = 7$.

As the expected cost of misclassification is lower when classifying the observations in class 2, all the observations in node 23 are classified as being in class 2.

We have obtained the full tree based on the given data. However, it is possible that this tree is over-fitting the data. Thus, we need to prune the tree to obtain the best possible tree to classify future observations (not only to classify the given data).

For any non-terminal nodes, that is, nodes 1, 2, 3, 5 and 11, we need to find the threshold at which the tree consisting of the given node only becomes better than the subtree having the given node as root node. Note that the cost of a node is the expected cost of misclassification (ECM) for this node.

We already know that ECM=3 for node 1. Moreover, the cost of the subtree having node 1 as root node (in fact, this subtree is the whole tree), $T_{\text{node 1}}$, is $(0 \times \frac{9}{40}) + (2.666 \times \frac{6}{40}) + (1.5 \times \frac{10}{40}) + (0 \times \frac{3}{40}) + (1.6 \times \frac{5}{40}) + (1.428571 \times \frac{7}{40}) = 1.225$. Thus, the threshold at which node 1 becomes a better subtree is

$$\lambda(1) = \frac{3 - 1.225}{6 - 1} = 0.355$$

as $| T_{\text{node 1}} | = 6$.

We do the same calculations for the 4 other non-terminal nodes. The costs of the trees consisting of node 2, 3, 5 and 11 only are respectively 2, 2.1875, 3 and 2.5. Moreover, the costs of the subtrees of these four nodes are respectively 0.75, 1.9375, 1.2 and 1.5. Thus, we can obtain the threshold for the values of the complexity parameter at which the trees consisting of one node only become better than their subtrees.

$$\lambda(2) = \frac{2 - 0.75}{4 - 1} = 0.4166$$

$$\lambda(3) = \frac{2.1875 - 1.9375}{2 - 1} = 0.25$$

$$\lambda(5) = \frac{3 - 1.2}{3 - 1} = 0.9$$

$$\lambda(11) = \frac{2.5 - 1.5}{2 - 1} = 1.$$

Among these 5 values, the smallest is $\lambda(3)$. Thus, we prune all the nodes below node 3, that is, nodes 6 and 7. The tree obtained is the tree T_2 and has associated complexity parameter $\lambda_2 = \lambda(3) = 0.25$. Based on T_2, we search for the subtree of this tree with the lower threshold for the complexity parameter. In T_2, node 3 is a terminal node, so it doesn't have a subtree. The costs of the subtrees of node 2, 5 and 11 don't change as these subtrees are still the same. However, the cost of the subtree of node

1 changes as we pruned 2 nodes from the tree. The cost of the subtree having node 1 as root node is $(0 \times \frac{9}{40}) + (2.1875 \times \frac{16}{40}) + (0 \times \frac{3}{40}) + (1.6 \times \frac{5}{40}) + (1.428571 \times \frac{7}{40}) = 1.325$, an increase compared to T_1 as we removed nodes from the tree. The cost of the tree consisting of the node 1 only remains 3, so

$$\lambda(1) = \frac{3 - 1.325}{5 - 1} = 0.41875.$$

Among the 3 thresholds, the smallest is $\lambda(2)$. Thus, we prune the nodes below node 2. The tree obtained, T_3, contains 3 nodes, nodes 1, 2 and 3, and has associated complexity parameter $\lambda_3 = \lambda(2) = 0.4166$.

Finally, in T_3, there is only 1 non-terminal node, node 1. The cost of the tree consisting of this node alone is still 3. The cost of the subtree of node 1, that is, the cost of the whole tree T_3, is $2 \times \frac{24}{40} + 2.1875 \times \frac{16}{40} = 2.0875$. Thus, the threshold at which node 1 becomes a better tree is

$$\lambda(1) = \frac{3 - 2.0875}{2 - 1} = 0.9125.$$

We prune nodes 2 and 3. We obtain the tree T_4, consisting only of node 1.

For the 4 subtrees obtained, we obtain the cross-validation estimate for the cost of the tree. Figure 2.1 shows a plot of the cross-validation estimate of the cost against the size (or complexity parameter) for each tree.

The tree with the lowest cross-validation estimate is T_2, where nodes 6 and 7 have been pruned off, giving a tree with 5 terminal nodes only. Thus, T_2 is the best pruned subtree of T and it is the tree that will be used to classify future observations.

The last thing we need to do is to determine in which class the observations in node 3, now a terminal node, must be classified. If we classify the observations in class 1, the ECM is

$$ECM_3 = c_{12} P_{Y|A}(2 \mid 3) = 8 \times \frac{9}{16} = 4.5$$

as $n_{23} = 9$ and $n_3 = 16$. Similarly, if we classify the observations in class 2, the ECM is

$$ECM_3 = c_{21} P_{Y|A}(1 \mid 3) = 5 \times \frac{7}{16} = 2.1875$$

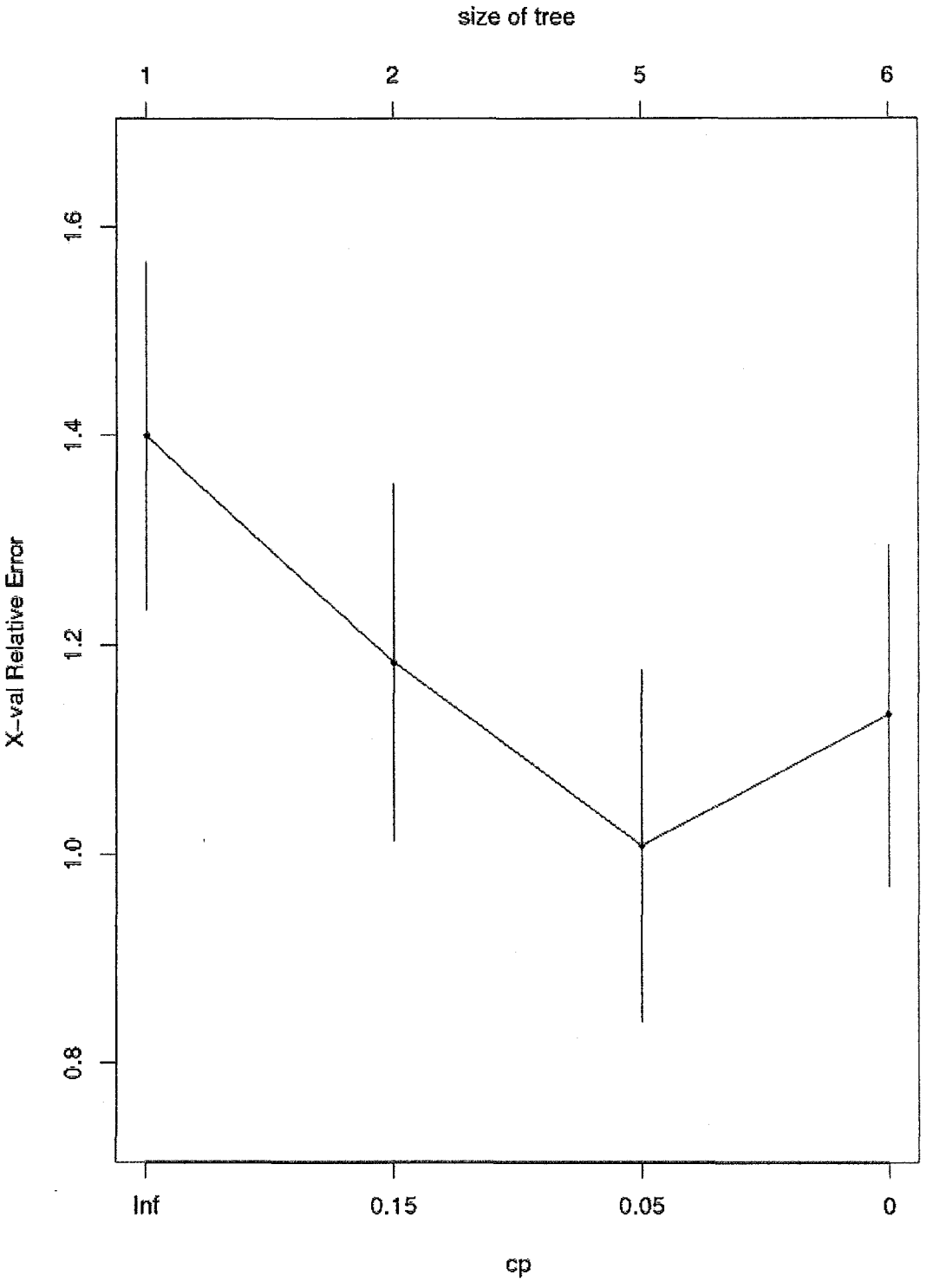

Figure 2.1: Cross-validation cost vs Size of the tree for example 2.6

as $n_{13} = 7$ and $n_3 = 16$.

As the expected cost of misclassification is lower when classifying the observations in class 2, all the observations in node 23 are classified as being in class 2.

The misclassification rate of this tree is 25% (10 of the 40 observations are misclassified), whereas it is 4% (1 of the 25 class 1 observations are misclassified) for class 1 observations only and 60% (9 of the 15 class 2 observations are misclassified) for class 2 observations. The final classification tree is shown in figure 2.2

2.7 Regression trees

There is another type of tree called a regression tree. A regression tree is used when the response variable is continuous. The most significant difference with a classification tree is that the splits must minimize the sum of squares of the differences between the true value of the response variable and the predicted value. For each split, a least square regression model is fitted to determine the variable on which to split and the split point. The split obtained minimizes the total sum of squares of errors of the two child nodes. The values predicted by the tree at the terminal nodes are the expected values of the response variable given the values of the predictor variables instead of classes. Otherwise, the two types of trees are very similar.

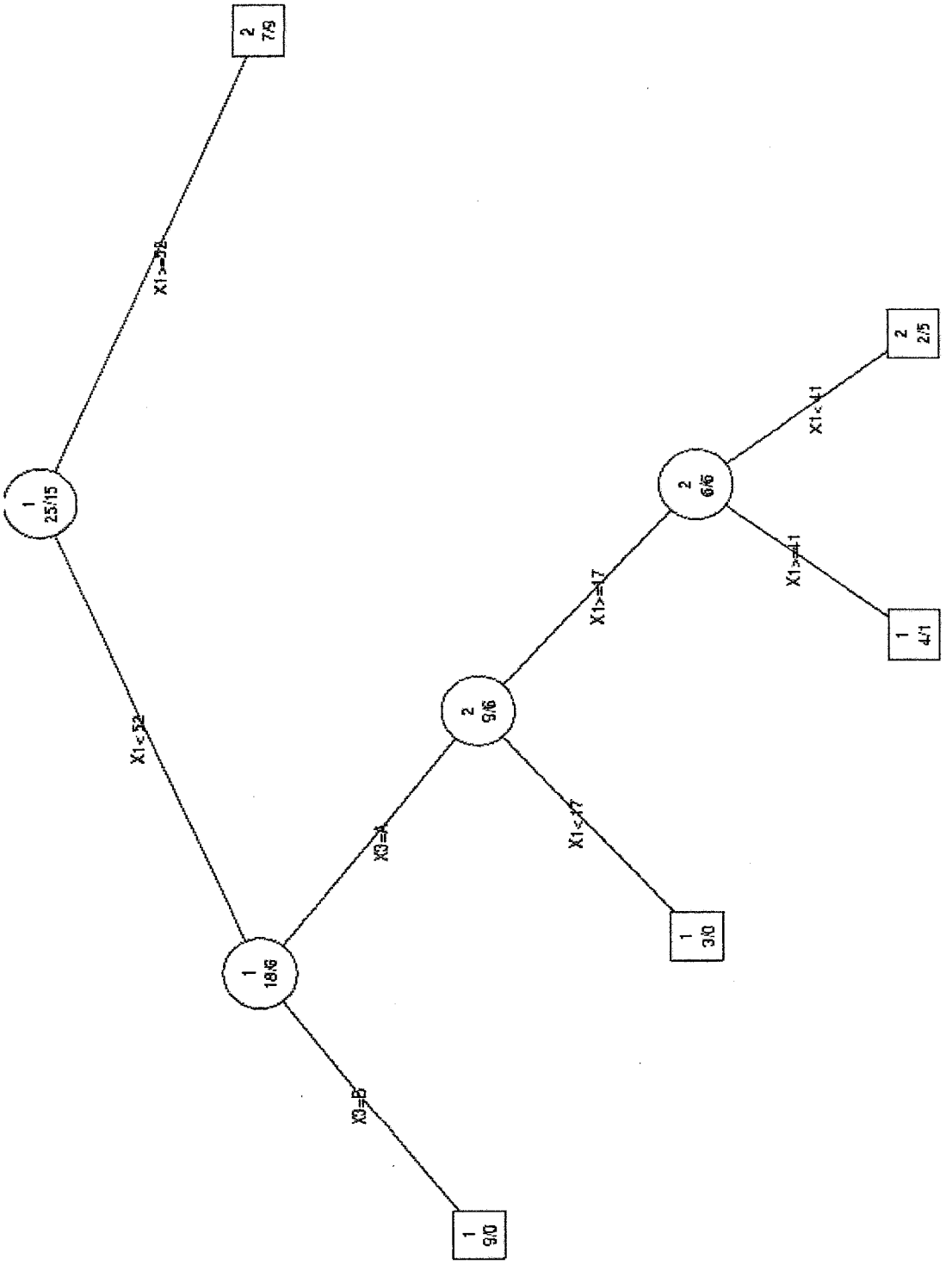

Figure 2.2: Best pruned tree for example 2.6

Chapter 3

Analysis of the NCDB data using classification trees

We now look at the results obtained when analyzing the NCDB data using classification trees. Two classification trees modeling respectively the risk of fatal collisions compared to non-fatal injury collisions for collisions involving drivers aged 65 years and older and the risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older have been grown. The selection of these trees is detailed in sections 3.1 and 3.2. Moreover, an interpretation of the resulting trees is given.

3.1 Modeling the risk of fatal collisions

Our first model analyzes the risk of fatal collisions compared to non-fatal injury collisions for collisions involving drivers aged 65 years and older. 78,893 collisions were selected for analysis.

Fatal collisions represent 2.0% (1,601 collisions) of the collisions analyzed whereas non-fatal injury collisions represent 98.0% (77,292) of the collisions. Note that as these empirical class probabilities are thought to be representative of the proportions of fatal collisions and non-fatal injury collisions in general, they were used as the prior class probabilities to grow the tree.

Two trees were first grown without using any misclassification costs, i.e. all misclassification costs are equal to 1: one using the Gini criterion as the impurity measure and the other using the deviance. Both trees classified all the observations as non-fatal injury collisions because fatal collisions comprised only 2% of all the collisions. Therefore, the total misclassification rates of both trees were very low. However, all the fatal collisions were misclassified whereas all the non-fatal injury collisions were classified correctly by these tree.

Misclassification costs were thus introduced in the model. As it is clearly more important to classify fatal collisions correctly than non-fatal injury collisions, the cost of misclassifying fatal collisions is higher than the cost misclassifying non-fatal collisions. Different combinations of costs were tried and the misclassification cost matrix was chosen such that fatal collisions had a low misclassification rate, but such that the misclassification rate for non-fatal injury collisions also remained below some reasonable limit. Table 3.1 shows the misclassification cost matrix chosen.

Table 3.1: Misclassification cost matrix for Model 1 and Model 2

	Severity of the collision	
Classification	**Non-fatal injury**	**Fatal**
Non-fatal injury	0	59
Fatal	1	0

Two new trees were grown (one using the Gini criterion as the impurity criterion and one using deviance) using this misclassification cost matrix. The best pruned trees are called Model 1 and Model 2 respectively. Table 3.2 summarizes the total misclassification rates, as well as the misclassification rates for fatal collisions and non-fatal injury collisions respectively for models 1 and 2.

Both trees have similar total misclassification rates, the misclassification rate for Model 2 being a little bit lower. Moreover, they both do much better than the trees grown using no misclassification costs. However, the tree produced using the Gini

Table 3.2: Misclassification rates for Model 1 and Model 2

Model	All collisions (%)	Fatal collisions (%)	Non-fatal injury collisions (%)
Model 1	31.9	14.4	32.3
Model 2	30.2	14.5	30.2

criterion contains many fewer nodes than the tree produced using deviance. As we want to obtain a model that is somewhat easy to interpret, we prefer a tree with fewer terminal nodes. Thus, Model 1 was chosen to model the risk of fatal collisions compared to non-fatal injury collisions among collisions involving drivers aged 65 years and older. As Table 3.2 shows, this tree is very good at classifying both fatal collisions and non-fatal injury collisions. The tree contains 24 terminal nodes and 10 levels. Figure 3.1 shows a plot of the cross-validation cost against the size of the tree for each possible choice of pruned tree. It shows that the best pruned tree should contain 24 terminal nodes.

Figure 3.2 shows the best pruned tree. The categories are numbered according to the ordering of the categories in the list of section 1.3.

The splits used in the tree show the factors having a significant influence when we are predicting the severity of a collision (fatal collision or non-fatal injury collision) for collisions involving drivers aged 65 years and older. Note that the higher a node is in the tree, the more important is the variable used in the split used to obtain that node. The variables retained in the best pruned tree, ordered according to their level in the tree, are:

- Road classification I
- Collision configuration
- Posted speed limit
- Number of unrestrained victims in the collisions involving the studied driver
- Number of restrained victims in the collision involving the studied driver

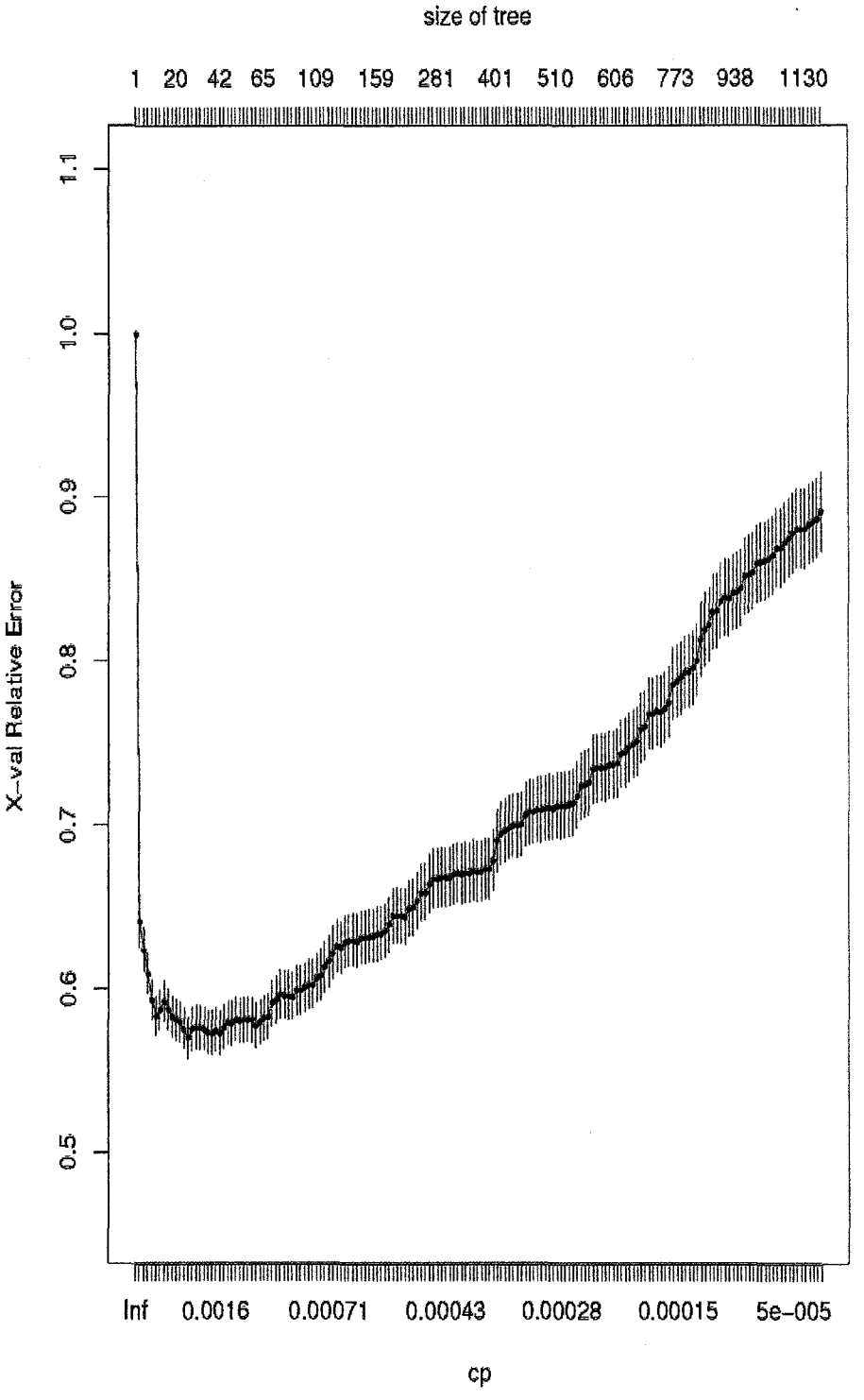

Figure 3.1: Cross-validation cost vs Size of the tree for Model 1

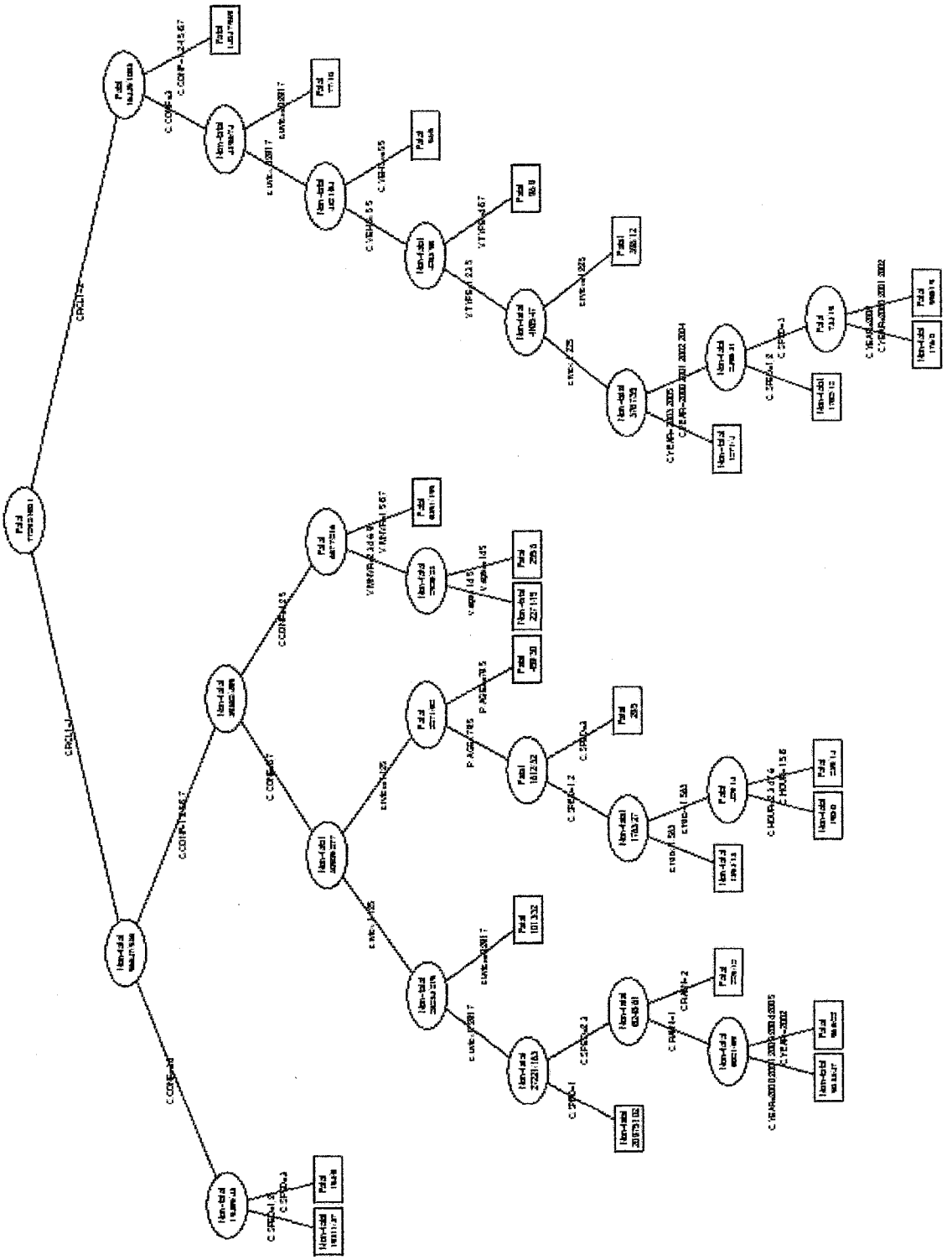

Figure 3.2: Best pruned tree for Model 1

- Vehicle manoeuvre
- Number of vehicles involved
- Driver age
- Vehicle age
- Vehicle type
- Road alignment
- Year of collision
- Hour of collision

A quick look at the tree tells us that the most important variable to predict the severity of a collision involving a driver aged 65 years and older is the type of road on which the collision occurred (rural or urban). The collision configuration is the second most important variables as both level 2 splits use this variable.

Observations in terminal nodes 1, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 19, 21 and 24 are classified as fatal collisions whereas observations in terminal nodes 2, 9, 12, 16, 17, 18, 20, 22 and 23 are classified as non-fatal injury collisions.

Terminal node 1 shows that all collisions on rural roads except rear-end collisions are more likely to be fatal collisions. This result is particularly significant as 989 of the 1,601 fatal collisions (62%) ended up in that node. Terminal node 3 shows that all collisions on urban roads where the posted speed limit is at least 100 km/h and involving vehicles going in the same direction are more likely to be fatal collisions. Terminal node 4 shows that rear-end collisions on rural roads where at least 1 victim was unrestrained are more likely to be fatal collisions. Terminal node 5 shows that single-vehicle collisions or head-on collisions on urban roads where the vehicle was going straight ahead, reversing, changing lanes or overtaking/passing prior to the collision are more likely to be fatal collisions. Terminal node 6 shows that rear-end collisions on rural roads involving 6 or more vehicles where the number of unrestrained injuries in the collision involving the studied driver is 0 are more likely to be fatal collisions.

Terminal node 7 shows that collisions on urban roads involving vehicles going in opposite directions except head-on collisions where the number of restrained victims in the collisions involving the studied drivers is at most 1 and the number of unrestrained victims in the collisions involving the studied drivers is at least 1 are more likely to be fatal collisions. Terminal node 8 shows that collisions on urban roads involving vehicles going in opposite directions except head-on collisions where the number of restrained victims in the collision involving the studied drivers is at least 2 and where the studied driver is at least 80 years old are more likely to be fatal collisions. Terminal node 10 shows that single-vehicle or head-on collisions on urban roads where the vehicle was turning left or right, making a U-turn, slowing or stopping in traffic or was performing any manoeuvre not listed in section 1.3 and where the vehicle was 15 years and older are more likely to be fatal collisions. Terminal node 11 shows that rear-end collisions on rural roads involving less than 6 vehicles where the number of unrestrained victims in the collision involving the studied driver is 0 and where the vehicle of the studied driver is an heavy truck, a motorcycle or an other vehicle (motorhome, etc.) are more likely to be fatal collisions.

Terminal node 13 shows that collisions on urban roads where the posted speed limit is at least 100 km/h involving two vehicles going in opposite direction except head-on collisions, where the number of restrained victims in the collision involving the studied driver is at least 2 and where the studied driver is between 65 and 79 years old are more likely to be fatal collisions. Terminal node 14 shows that rear-end collisions on rural roads involving less than 6 vehicles where the number of unrestrained victims in the collision involving the studied driver is 0 and the number of restrained victims in the collision involving the studied driver is at least 2 and where the vehicle of the studied driver is a passenger car, a passenger van or SUV, a light truck or a bus are more likely to be fatal collisions.

Terminal node 15 shows that collisions on urban roads involving two vehicles going in opposite direction except head-on collisions where the posted speed limit is

at least 60 km/h, where the road is curved at the collision site and where the number of restrained victims in the collision involving the studied driver is at most 1 and the number of unrestrained victims in the collision involving the studied driver is 0 are more likely to be fatal collisions. Terminal node 19 shows that collisions which occurred in 2002 on urban roads involving two vehicles going in opposite direction except head-on collisions where the posted speed limit is at least 60 km/h, where the road is going straight at the collision site and where the number of restrained victims in the collision involving the studied driver is at most 1 and the number of unrestrained victims in the collision involving the studied driver is 0 are more likely to be fatal collisions.

Terminal node 21 shows that collisions on urban roads where the posted speed limit is at most 90 km/h involving two vehicles going in opposite direction except head-on collisions, where the number of restrained victims in the collision involving the studied driver is at least 2 and where the studied driver is between 65 and 79 years old and which occurred between 0:00 and 3:00 or between 12:00 and 18:00 are more likely to be fatal collisions. Finally, terminal node 24 shows that rear-end collisions on rural roads where the posted speed limit is at least 100 km/h involving less than 6 vehicles where the number of unrestrained victims in the collision involving the studied driver is 0 and the number of restrained victims in the collision involving the studied driver is at most 1, where the vehicle of the studied driver is a passenger car, a passenger van or SUV, a light truck or a bus and which occurred between 2000 and 2002 are more likely to be fatal collisions.

3.2 Modeling the risk of fatal injuries to the driver

The risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older is now modeled. 40,221 drivers were selected for analysis. Fatally injured drivers represent 2.2% (903 drivers) of the drivers analyzed whereas non-fatally in-

jured drivers represent 97.8% (39,318) of the drivers. Note that as these empirical class probabilities are thought to be representative of the proportions of fatally injured drivers and non-fatally injured drivers in general, they were used as the prior class probabilities to grow the tree.

Two trees were grown without using any misclassification costs, i.e. all misclassification costs are equal to 1: one using the Gini criterion as the impurity measure and the other using deviance. Both trees classified all the observations as non-fatally injured drivers because fatally injured drivers comprised only 2.2% of all the drivers. As previously observed, although the total misclassification rate is low for these trees, it is because fatally injured drivers comprise only 2.2% of all the drivers. Misclassification costs were thus introduced in the model. Table 3.3 shows the misclassification cost matrix chosen.

Table 3.3: Misclassification cost matrix for Model 3 and Model 4

	Severity of the injuries to the driver	
Classification	**Non-fatally injured**	**Fatally injured**
Non-fatally injured	0	46
Fatally injured	1	0

Two new trees were grown (one using the Gini criterion as the impurity criterion and one using deviance) using this misclassification cost matrix. The best pruned trees using each criterion are called Model 3 and Model 4 respectively. Table 3.4 summarizes the total misclassification rates, as well as the misclassification rates for fatally injured drivers and non-fatally injured drivers respectively for models 3 and 4.

Both trees have somewhat similar total misclassification rate for this data set, the misclassification rate for model 3 being a little bit lower. Moreover, they both do much better than the trees grown using no misclassification costs. However, once again, the tree produced using the Gini criterion contains many fewer terminal nodes and it was thus preferred to the other tree. Thus, Model 3 was chosen to model the

Table 3.4: Misclassification rates for Model 3 and Model 4

Model	All drivers (%)	Fatally injured drivers (%)	Non-fatally injured drivers (%)
Model 3	32.4	14.4	32.9
Model 4	27.4	14.7	27.8

risk of fatal injuries for drivers aged 65 years and older. As Table 3.4 shows, this tree is very good at classifying both fatally injured drivers and non-fatally injured drivers. The best pruned tree contains 25 terminal nodes and 12 levels. Figure 3.3 shows a plot of the cross-validation cost against the size of the tree for each possible choice of pruned tree. It shows that the best pruned tree should contain 25 terminal nodes. Figure 3.4 shows the best pruned tree.

The splits used in the tree show which factors have a significant influence on the severity of the injuries for drivers aged 65 years and older (fatal injuries or non-fatal injuries). The variables retained in the best pruned tree, ordered according to their level in the tree, are:

- Road classification I
- Collision configuration
- Safety device used
- Driver age
- Driver sex
- Weather condition
- Improper turning
- Other driver condition
- Month of collision
- Traffic control
- Vehicle manoeuvre
- Hour of collision

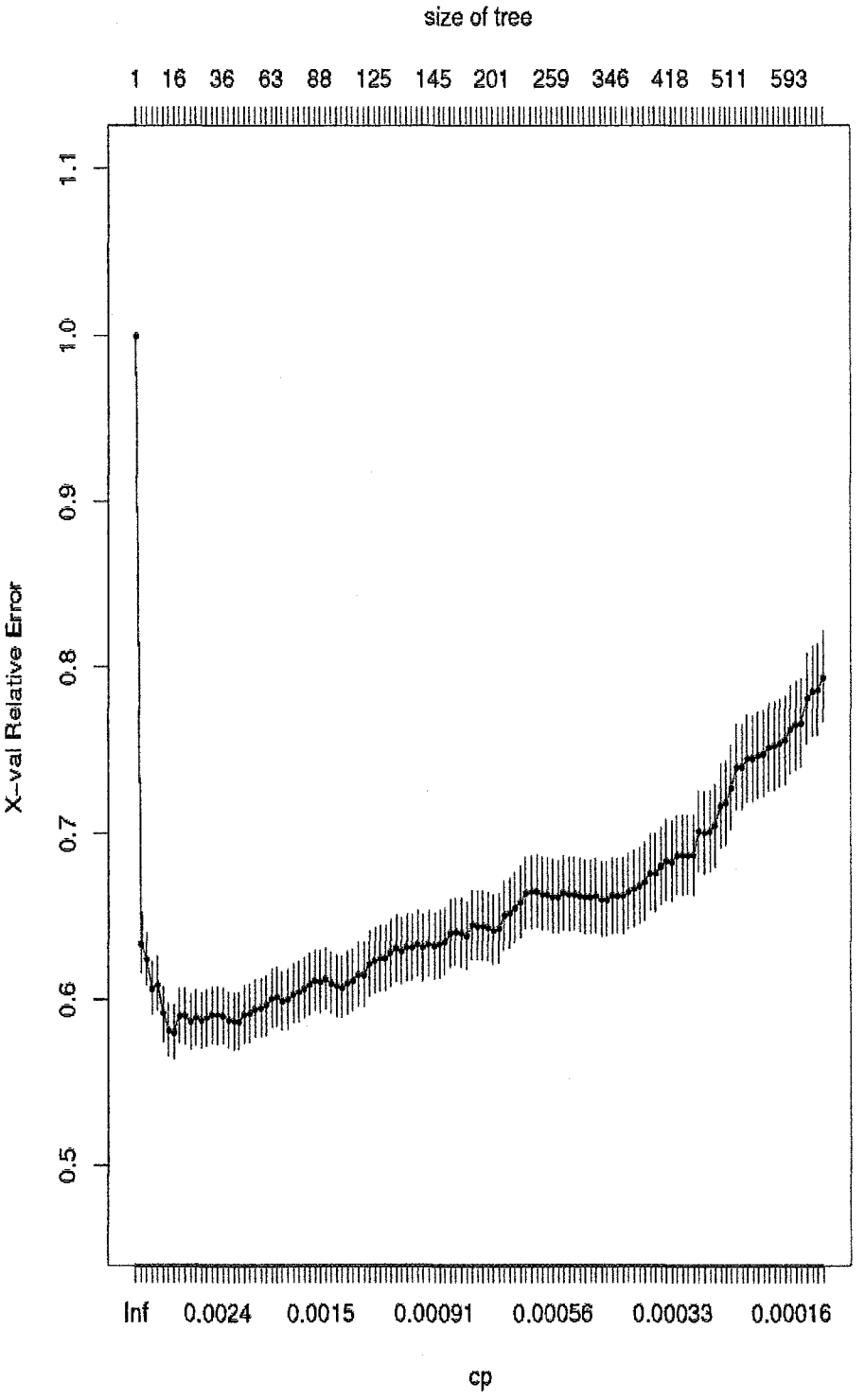

Figure 3.3: Cross-validation cost vs Size of the tree for Model 3

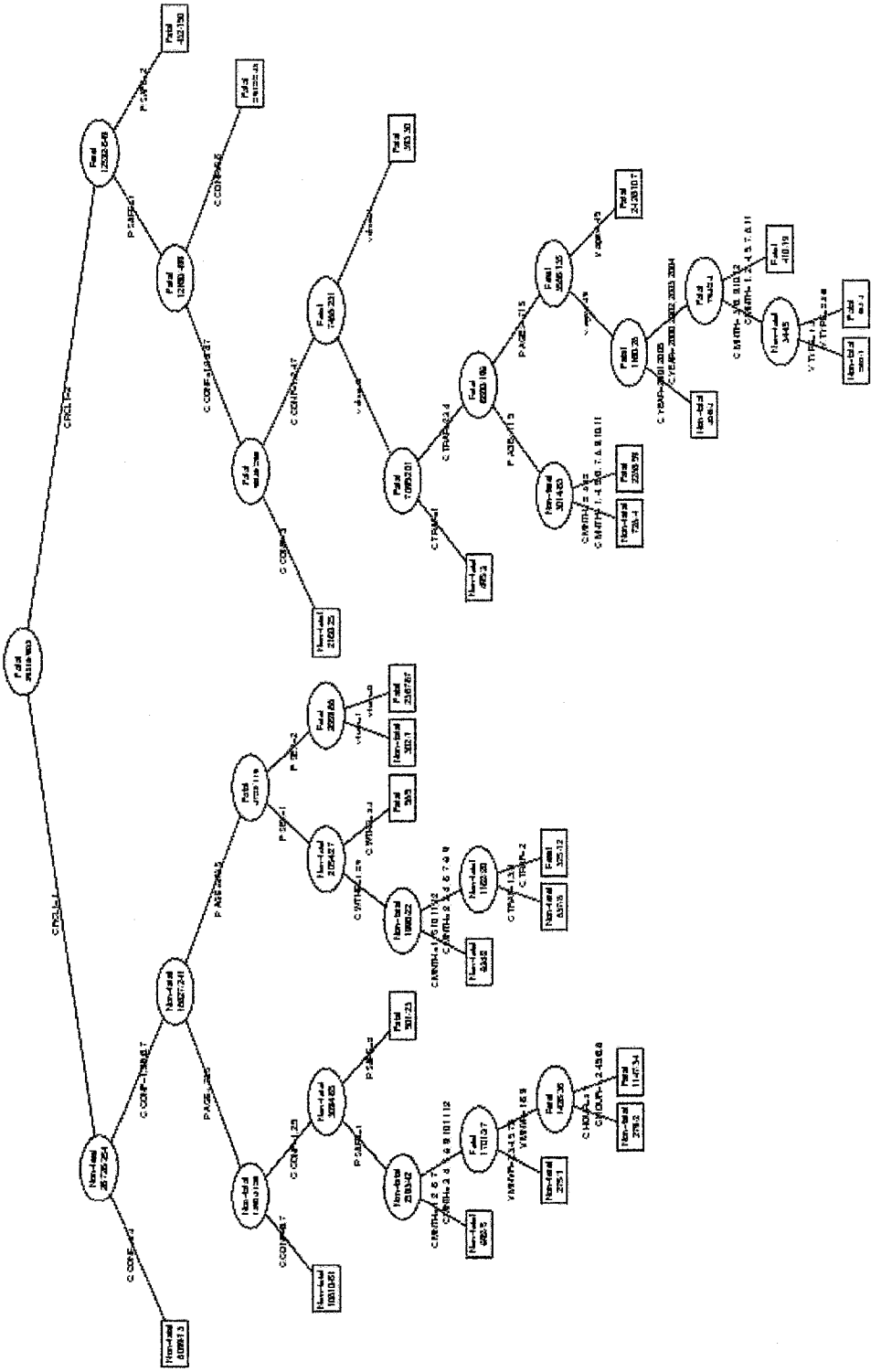

Figure 3.4: Best pruned tree for Model 3

- Vehicle age
- Year of collision
- Vehicle type.

A quick look at the tree tells us that the most important variable to predict the severity of the injuries to drivers aged 65 years and older is again the type of road on which the collision occurred (rural or urban). Collision configuration and Safety device used are the second most important variables (level 2 splits).

Observations in terminal nodes 2, 3, 6, 7, 9, 10, 16, 18, 20, 21, 23 and 25 are classified as fatally injured drivers whereas observations in terminal nodes 1, 4, 5, 8, 11, 12, 13, 14, 15, 17, 19, 22 and 24 are classified as non-fatally injured drivers.

Terminal node 2 shows that unrestrained drivers involved in collisions on rural roads are more likely to be fatally injured. Terminal node 3 shows that restrained drivers involved in head-on or right angle collisions on rural roads are more likely to be fatally injured. Terminal node 6 shows that unrestrained drivers aged between 65 and 78 years old involved in single-vehicle collisions or head-on collisions on urban roads are more likely to be fatally injured. Terminal node 7 shows that female drivers aged 79 years and older involved in single-vehicle collisions or two-vehicle collisions on urban roads where vehicles were going in opposite direction and which occurred while it was snowing or falling freezing rain, sleet or hail or while the visibility was limited are more likely to be fatally injured. Terminal node 9 shows that male drivers aged 79 years and older involved in single-vehicle collisions or two-vehicle collisions on urban roads where vehicles were going in opposite directions and who didn't do an improper turn are more likely to be fatally injured.

Terminal node 10 shows that restrained drivers involved in single-vehicle collisions or two-vehicle collisions other than rear-end, head-on or right angle collisions, on rural roads and which suffered a condition other than those mentioned in the list in section 1.3 are more likely to be fatally injured. Terminal node 16 shows that male drivers aged 79 years and older involved in single-vehicle collisions or two-vehicle col-

lisions on urban roads where the vehicles were going in opposite directions, where signs were used to control traffic at the collision site, where it was raining, there was strong winds or it was clear and sunny at the time of the collision and which occurred during the months of February to April or June to September are more likely to be fatally injured. Terminal node 18 shows that restrained drivers aged between 65 and 78 years old involved in single-vehicle collisions or head-on collisions on rural roads where the vehicle was going straight ahead, reversing or performing another manoeuvre not mentioned in the list prior to the collision and which occurred during the months of March to May or August to December between 21:00 and 6:00 or 9:00 and 18:00 are more likely to be fatally injured.

Terminal node 20 shows that restrained drivers aged between 65 and 71 years old which didn't suffer a condition other than those mentioned in the list in section 1.3 involved in single-vehicle collisions or two-vehicle collisions other than rear-end, head-on or right-angle collisions, on rural roads where no traffic control device or traffic control devices other than traffic lights were present at the collision site and which occurred during the months of January or April to December are more likely to be fatally injured. Terminal node 21 shows that restrained drivers 72 years and older which didn't suffer a condition other than those mentioned in the list in section 1.3 involved in single-vehicle collisions or two-vehicle collisions other than rear-end, head-on or right-angle collisions, on rural roads where no traffic control device or traffic control devices other than traffic lights were present at the collision site and where the vehicle of the studied driver is 5 years or older are more likely to be fatally injured.

Terminal node 23 shows that restrained drivers 72 years and older which didn't suffer a condition other than those mentioned in the list in section 1.3 involved in single-vehicle collisions or two-vehicle collisions other than rear-end, head-on or right-angle collisions, on rural roads where no traffic control device or traffic control devices other than traffic lights were present at the collision site, where the vehicle of the

studied driver is 4 years old or less and which occurred during the months of January, February, April, May, July, August or November in 2000, 2002, 2003 or 2004 are more likely to be fatally injured. Finally, terminal node 25 shows that restrained drivers 72 years and older which didn't suffer a condition other than those mentioned in the list in section 1.3 involved in single-vehicle collisions or two-vehicle collisions other than rear-end, head-on or right-angle collisions, on rural roads where the vehicle of the studied driver is not a passenger car or a heavy truck, where no traffic control device or traffic control devices other than traffic lights were present at the collision site, where the vehicle of the studied driver is 4 years old or less and which occurred during the months of March, June, September, October or December in 2000, 2002, 2003 or 2004 are more likely to be fatally injured.

Chapter 4

Logistic regression

In the problem addressed in this thesis, the response variable Y is a categorical variable. In this case, linear regression or other types of regression models that were developed to model relationships where the response variable was continuous, are not useful to fit the data. The logistic regression model is a better fit for such data and is widely used to model the relationship between this kind of response variable and a set of predictor variables. Note that the material in this section is based on the material from Agresti (2002), McCullagh and Nelder (1989) and Neter, Kutner, Nachsteim and Wasserman(1996) [11, 12, 13]

4.1 Generalized linear models

Generalized linear models (GLM) are regression models that can be made linear by applying a given transformation to the response variable Y. These models consist of three major components : a response variable, a linear predictor function and a link function.

The first component, the response variable Y, should have a distribution from an exponential family. The second component of a GLM is a linear predictor function. A linear predictor function is a linear combination of the predictor variables used in

the model. That is, the linear predictor should be of the form

$$\sum_{j=1}^{k} \beta_j x_{ij}, i = 1, \ldots, n,$$

where x_{ij} is the value of the j-th predictor variable for observation i and β_j is a parameter representing the effect of the j-th predictor variable on Y. The third component, the link function, is the transformation g applied to the response variable to make the model linear. That is, g must be a function such that $g(E(Y)) = \sum_{j=1}^{k} \beta_j x_{ij}$, which is linear.

A GLM is thus of the form

$$g(E(Y)) = \sum_{j=1}^{k} \beta_j x_{ij}, i = 1, \ldots, n,$$

Several known models are generalized linear models. This the case of the linear regression model. For this model, the link function, also called the identity link, is given by $g(x) = x$. The binary logistic regression model (also called the logit model), which will be described in more detail in section 4.2, is also a GLM.

4.2 Binary logistic regression

Let Y be a binary response variable and let $\mathbf{x} = (x_1, \ldots, x_k)$ be a vector containing the values of k predictor variables associated with Y, which can be either discrete, continuous or categorical. As Y is a binary variable, let Y have a Bernoulli distribution with probability $\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x})$ of success and probability $1 - \pi(\mathbf{x}) = P(Y = 0 \mid \mathbf{x})$ of failure.

Define the odds of a success as the probability of a success over the probability of a failure, that is,

$$Odds = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}.$$

The logit of π is then defined as the natural logarithm of the odds of success:

$$logit(\pi(\mathbf{x})) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right)$$

Thus, the binary logistic regression model is given by

$$logit(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

where $\pi(\mathbf{x}) = P(Y = 1 \mid X_1 = x_1, \ldots, X_k = x_k)$. The logit is the link function used in a binary logistic regression model.

In matrix notation, the binary logistic regression model is written as

$$logit(\pi(\mathbf{x})) = \beta'\mathbf{x}$$

where $\beta' = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{bmatrix}$.

The binary logistic regression model can also be written in terms of the probability of success instead of in terms of the logit. The model would thus have the form:

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}},$$

or in matrix notation

$$\pi(\mathbf{x}) = \frac{e^{\beta'\mathbf{x}}}{1 + e^{\beta'\mathbf{x}}}.$$

4.3 Fitting the logistic regression model

As the parameters $\beta_j, j = 1, \ldots, k$ are unknown in such a model, we need to estimate their values by fitting the model to the data. To achieve this, the maximum likelihood estimators of the parameters are obtained.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be random variables representing the value of the response variable Y for a sample of n observations. Y_i is taken from a $Bernoulli(\pi_i = \pi(\mathbf{x_i}))$ distribution for $i = 1, \ldots, n$. Moreover, let the matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1k} & \ldots & x_{nk} \end{bmatrix}$$

be a matrix containing the values of the k predictor variables for each of these n observations. Note that these predictor variables can be either discrete, continuous or categorical.

As $Y_1, \ldots, Y_n$ have a $Bernoulli(\pi_i)$, $i = 1, \ldots, n$ distribution, then:

$$f_{Y_i}(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where $y_i = 0$ or $1, i = 1, \ldots, n$. As $Y_1, \ldots, Y_n$ are independent observations, the joint probability density function of $Y_1, \ldots, Y_n$, called the likelihood function, is

$$L(\beta) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

The log-likelihood function is then obtained by taking the natural logarithm of this function, as it is easier to work with such a function. The log-likelihood function is given by

$$\begin{aligned}
\ln L(\beta) &= \ln f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) \\
&= \ln \prod_{i=1}^{n} f_{Y_i}(y_i) \\
&= \ln \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \sum_{i=1}^{n} (\ln \pi_i^{y_i} + \ln(1-\pi_i)^{1-y_i}) \\
&= \sum_{i=1}^{n} (y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)) \\
&= \sum_{i=1}^{n} (y_i \ln \pi_i + \ln(1-\pi_i) - y_i \ln(1-\pi_i)) \\
&= \sum_{i=1}^{n} (y_i \ln \left(\frac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)) \\
&= \sum_{i=1}^{n} y_i \ln \left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i).
\end{aligned}$$

However, remember that $\pi_i = E(Y_i) = \frac{e^{\beta' \mathbf{X_i}}}{1+e^{\beta' \mathbf{X_i}}}$ where $\mathbf{X_i}$ is the i-th column of the

matrix $\mathbf{X}$. Thus,

$$\begin{aligned}
\ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \ln \frac{\frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}}{1-\frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}} \\
&= \ln \frac{\frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}}{\frac{1+\exp\beta'\mathbf{X_i}-\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}} \\
&= \ln \exp\beta'\mathbf{X_i} \\
&= \beta'\mathbf{X_i}.
\end{aligned}$$

So, the log-likelihood function is given by

$$\begin{aligned}
\ln L(\beta) &= \ln f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) \\
&= \sum_{i=1}^{n} y_i\beta'\mathbf{X_i} + \sum_{i=1}^{n} \ln\left(1-\frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}\right) \\
&= \sum_{i=1}^{n} y_i\beta'\mathbf{X_i} + \sum_{i=1}^{n} \ln\left(\frac{1}{1+\exp\beta'\mathbf{X_i}}\right) \\
&= \sum_{i=1}^{n} y_i\beta'\mathbf{X_i} + \sum_{i=1}^{n} \ln(1+\exp\beta'\mathbf{X_i})^{-1} \\
&= \sum_{i=1}^{n} y_i\beta'\mathbf{X_i} - \sum_{i=1}^{n} \ln(1+\exp\beta'\mathbf{X_i}).
\end{aligned}$$

The maximum-likelihood estimator of the vector β is then obtained by finding the maximum of the log-likelihood function (as maximizing the log-likelihood function is the same as maximizing the likelihood function). This is achieved by finding the partial derivatives of $\ln L(\beta)$ with respect to $\beta_0,\ldots,\beta_k$ and equating them to 0. Then, we solve for $\beta_0,\ldots,\beta_k$. The partial derivatives of the log-likelihood function are given by

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i X_{ij} - \sum_{i=1}^{n} X_{ij}\frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}, j = 1,\ldots,k$$

However, as $\pi_i = \frac{\exp\beta'\mathbf{X_i}}{1+\exp\beta'\mathbf{X_i}}$,

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i X_{ij} - \sum_{i=1}^{n} \pi_i X_{ij}.$$

Then, we equate the partial derivatives to 0 and solve for $\beta_0, \ldots, \beta_k$. Thus, for $j = 1, \ldots, k$,

$$\sum_{i=1}^{n} y_i X_{ij} - \sum_{i=1}^{n} \pi_i X_{ij} = 0$$

However, these equations are not linear in terms of $\beta_j, j = 1, \ldots, k$. So, there is no closed form for $\beta_0, \ldots, \beta_k$. Thus, numerical methods need to be used to estimate the parameters. The Newton-Raphson method or Fisher's scoring, applied to logistic regression, are among the methods which can be used.

The Newton-Raphson method works as follows. Let

$$\mathbf{u} = \begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_k} \end{bmatrix}$$

be the vector of the partial derivatives of β and let

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 L(\beta)}{\partial \beta_1^2} & \frac{\partial^2 L(\beta)}{\partial \beta_1 \beta_2} & \cdots & \frac{\partial^2 L(\beta)}{\partial \beta_1 \beta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 L(\beta)}{\partial \beta_k \beta_1} & \frac{\partial^2 L(\beta)}{\partial \beta_k \beta_2} & \cdots & \frac{\partial^2 L(\beta)}{\partial \beta_k^2} \end{bmatrix}$$

be the vector of the second partial derivatives of β, also called the Hessian matrix.

Let also $\beta^{(0)}$ be the initial guess for the vector of parameters β and let $\mathbf{u}^{(0)}$ and $\mathbf{H}^{(0)}$ be $\mathbf{u}$ and $\mathbf{H}$ evaluated at the initial guess for β. We approximate $L(\beta)$ by its second-degree Taylor series expansion near $\beta^{(0)}$, given by

$$L(\beta) \approx L(\beta^{(0)}) + \mathbf{u}^{(0)\prime}(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})'\mathbf{H}^{(0)}(\beta - \beta^{(0)}).$$

Then, we find the vector of partial derivatives of $L(\beta)$, given by:

$$\frac{\partial L(\beta)}{\partial \beta} \approx \mathbf{u}^{(0)} + \mathbf{H}^{(0)}(\beta - \beta^{(0)}).$$

We equate this vector to 0 and solve for the next guess $\beta^{(1)}$. We get:

$$\mathbf{u}^{(0)} + \mathbf{H}^{(0)}(\beta - \beta^{(0)}) = 0$$

$$\Rightarrow \quad \mathbf{H}^{(0)}(\beta - \beta^{(0)}) = -\mathbf{u}^{(0)}$$

$$\Rightarrow \quad (\mathbf{H}^{(0)})^{-1}\mathbf{H}^{(0)}(\beta - \beta^{(0)}) = -(\mathbf{H}^{(0)})^{-1}\mathbf{u}^{(0)}$$

$$\Rightarrow \quad \beta - \beta^{(0)} = -(\mathbf{H}^{(0)})^{-1}\mathbf{u}^{(0)}$$

$$\Rightarrow \quad \beta = \beta^{(0)} - (\mathbf{H}^{(0)})^{-1}\mathbf{u}^{(0)}.$$

Thus, guess 1 is given by

$$\beta^{(1)} = \beta^{(0)} - (\mathbf{H}^{(0)})^{-1}\mathbf{u}^{(0)}$$

given that $\mathbf{H}^{(0)}$ has an inverse. We repeat this process in an iterative manner until the values of the guesses converge to some fixed vector. Thus, the guess $t+1$ for β, $\beta^{(t+1)}$ is given by:

$$\beta^{(t+1)} = \beta^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)}$$

where $\mathbf{u}^{(t)}$ and $\mathbf{H}^{(t)}$ are $\mathbf{u}$ and $\mathbf{H}$ evaluated at guess t for β.

The vector to which the guesses converge is an estimate of the maximum-likelihood estimator of the vector of parameters β.

Fisher's scoring works similarly to Newton-Raphson method. The only difference between the two methods is that the Hessian matrix $\mathbf{H}$ of β is replaced by $-E(\mathbf{H})$. Thus, the guess t for the vector β becomes

$$\beta^{(t+1)} = \beta^{(t)} + (E(\mathbf{H}^{(t)}))^{-1}\mathbf{u}^{(t)}.$$

Otherwise, both methods work exactly the same way. Fisher's scoring is the method used by SAS.

4.4 Categorical predictor variables

The predictor variables in binary logistic regression models can, among other things, be categorical variables. However, these variables are usually replaced by a set of binary dummy variables, as their use makes the model easier to interpret.

Suppose a categorical predictor variable has L categories. Define the variables $W_l, l = 1, \ldots, L$, as

$$W_l = \begin{cases} 1 & \text{if the observation is in category } \ell \\ 0 & \text{otherwise} \end{cases}$$

Only one of these variables is different from 0 for each observation as each observation can be in only one of the categories. Thus, $W_1 + \ldots + W_L = 1$ or $W_L = 1 - W_1 - \cdots - W_{L-1}$. So, only $W_1, \ldots, W_{L-1}$ are used as W_L is redundant. Category L is called the baseline category. As the information contained in all these variables is the same as that contained in the original categorical predictor variable, these variables can be used to replace the original ones.

4.5 Interpretation of the model

An important part of the use of any statistical model is to interpret the fitted model. In a binary logistic regression model, the parameter β_0 doesn't provide much information. However, the parameters $\beta_j, j = 1, \ldots, k$ are very important in the interpretation of the model. $\beta_j, j = 1, \ldots, k$, represents the effect of the variable X_j on the value of the logit. This effect can be interpreted as follows: for each increase of one unit of the value of the variable X_j, the logit increases (if $\beta_j \geq 0$) or decreases (if $\beta_i \leq 0$) by a value of β_j.

Translating the interpretation in terms of the odds of success, we say that each time the value of the variable X_j increases by one unit, the odds of success are multiplied by e^{β_j}. This value is called the odds ratio for the effect β_j as it can be obtained by dividing the odds of success when $X_j = x_j + 1$ by the odds of success when $X_j = x_j$, all the other variables remaining constant.

Confidence intervals for the logit and the odds of success can also be obtained to facilitate the interpretation of the model. A $(1 - \alpha)100\%$ confidence interval for the

logit of π_i is given by:

$$\widehat{\beta}'\mathbf{x} \pm z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})$$

where $SE(\widehat{\beta}'\mathbf{x})$ is the standard error of $\widehat{\beta}'\mathbf{x}$.

Exponentiating this confidence interval gives a $(1-\alpha)100\%$ confidence interval for the odds of success:

$$e^{\widehat{\beta}'\mathbf{x} \pm z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})}.$$

A $(1-\alpha)100\%$ confidence interval for the probability of success can also be found. It is given by:

$$\left[\frac{e^{\widehat{\beta}'\mathbf{x} - z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})}}{1 - e^{\widehat{\beta}'\mathbf{x} - z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})}}, \frac{e^{\widehat{\beta}'\mathbf{x} + z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})}}{1 - e^{\widehat{\beta}'\mathbf{x} + z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}'\mathbf{x})}} \right].$$

Finally, $(1-\alpha)100\%$ confidence intervals for the individual parameters β_j, $j = 1, \ldots, k$ can be obtained. They are given by

$$\widehat{\beta}_j \pm z_{\frac{\alpha}{2}} \times SE(\widehat{\beta}_j), j = 1, \ldots, k$$

where $SE(\widehat{\beta}_j)$ is the standard error of $\widehat{\beta}_j$.

$(1-\alpha)100\%$ confidence intervals for the odds ratio of the effect β_j, $j = 1, \ldots, k$ are obtained by exponentiating the confidence interval for β_j. They are given by:

$$e^{\widehat{\beta_j} \pm z_{\frac{\alpha}{2}} \times SE(\widehat{\beta_j})}, j = 1, \ldots, k.$$

4.6 Testing the model

Checking how well a model fits the data is another important step to perform when using logistic regression. One of the things that can be done to assess the fit is to check the significance of each individual predictor variable in the model. That is, we would like to test the following hypotheses:

$$H_0 : \beta_j = 0 \text{ vs } H_a : \beta_j \neq 0$$

for $j = 1, \ldots, k$. Wald's test or the likelihood-ratio test are usually used to test such hypotheses.

Wald's test uses the test statistic

$$Z = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$$

where $\widehat{\beta}_j$ is the estimated effect of variable X_j obtained by fitting the model and $SE(\widehat{\beta}_j)$ is the standard error of $\widehat{\beta}_j$. Usually, Z^2, distributed as a $\chi^2(1)$ is used as the test statistic. H_0 is rejected at level α if $Z^2 \geq \chi^2_\alpha(1)$ or if $p - value \leq \alpha$. That is, β_j is significant and the variable X_j should be kept in the model.

The likelihood-ratio test is also sometimes used. Let L_0 be the log-likelihood of the model under $\widehat{\beta}_j = 0$ and L_1 be the log-likelihood for the full model (that is, when $\widehat{\beta}_j \neq 0$). Then, the likelihood-ratio test uses the test statistic

$$R = -2(L_0 - L_1)$$

R, as Z^2, follows a $\chi^2(1)$ distribution. Thus, similar to Wald's test, H_0 is rejected at level α if $R \geq \chi^2_\alpha(1)$ or $p - value \leq \alpha$.

However, the most important thing to do is to test the model as a whole by comparing it to a more complex model. The likelihood-ratio test statistic G^2 or the Pearson χ^2 statistic, X^2, can be used to perform this test.

Suppose we would like to compare two models M_0 and M_1. The likelihood-ratio test statistic for comparing these two models would be given by

$$G^2(M_0 \mid M_1) = G^2(M_0) - G^2(M_1)$$

where $G^2(M_0)$ is the log-likelihood for the simpler model and $G^2(M_1)$ is the log-likelihood for the more complex model.

G^2 follows a $\chi^2(df)$ distribution where df = number of parameters in M_1 - number of parameters in M_0. We reject the hypothesis that the model fit is good at level α if $G^2 \geq \chi^2_\alpha(df)$.

Pearson's χ^2 statistic compares the observed frequency of each value of the vector of predictor variables and the fitted (or expected) frequency for each of these

values. However, Pearson's χ^2 statistic works for categorical data only, and therefore numerical data should be grouped into categories to use Pearson's χ^2 statistic.

Suppose M possible values are possible for the vector of predictor variables in a given model. Thus, Pearson's χ^2 statistic is given by:

$$X^2 = \sum_{m=1}^{M} \frac{(O_m - E_m)^2}{E_m}.$$

where O_m represents observed frequency of the m-th possible value and E_m represents the expected frequency of the m-th possible value. The expected frequency of the m-th value is calculated by multiplying the fitted probability of the m-th value by the total number of observations in the data. That is,

$$E_m = n\widehat{\pi_m}$$

X^2 follows a $\chi^2(M-p)$ distribution where p=number of parameters in the model. We reject the hypothesis that the model fit is good at level α if $X^2 \geq \chi^2_\alpha(M-p)$.

We may also want to test if a particular set of β_j's are equal to 0. We usually use the likelihood-ratio test. In that case, M_0 is the model where all these β_j's are equal to 0 and M_1 is the model where they are not 0.

4.7 Model selection procedures

There is no simple way to select a model and to determine which variables should be kept in the model or removed. For example, one may want to choose the variables in the model for some theoretical reason related to the subject studied, even though they are not all significant. However, if one really wants to find one of the best possible models in terms of statistical significance, some model selection procedures are available. However, none of these methods guarantees to give exactly the best possible model. Thus, one should be careful while using such procedures. Variables that are thought to be important for some reason should be kept in the model. Two of

the most well known model selection procedures are forward selection and backward elimination.

In forward selection, we begin with a model containing only the intercept term and then one variable is added at each step based on the improvement of the model fit when adding the chosen variable. The variable which most improves the fit of the model is added at each step. This is done until adding a new variable doesn't significantly improve the model fit. A more complex version of this technique tests after each step if the variables that were already in the model are still significant after adding the last variable.

Backward elimination works in a similar way, except it starts with the full model and removes variables for which the deletion is the least damaging for the model fit, that is the variable with the highest p-value in the model at this time. The process is stopped when deleting a new term gives a significantly poorer fit.

Chapter 5

Analysis of the NCDB data using logistic regression

We now analyze the NCDB data using logistic regression. Two logistic regression models are used to model respectively the risk of fatal collisions compared to non-fatal injury collisions for collisions involving drivers aged 65 years and older and the risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older. These models are the logistic regression equivalent of the classification tree models obtained in Chapter 3. The two fitted models are detailed in sections 5.1 and 5.2 respectively. Moreover, an interpretation of these models is given. Based on the fitted logistic regression models, odds ratios and their 95% confidence intervals were obtained for each effect. The p-values for the significance of each effect were also obtained. The logistic regression models were fitted using only the cases with no missing values.

5.1 Modeling the risk of fatal collisions

We first modeled the risk of fatal collisions compared to non-fatal injury collisions for collisions involving drivers aged 65 years and older. As only the cases without missing values are used, 47,965 collisions were selected for analysis. 857 fatal collisions and

47,108 non-fatal injury collisions were included in the analysis. A preliminary model using all the predictor variables has first been fitted. The model obtained is called Model 1. The fitted model was tested for goodness-of-fit using the likelihood-ratio test. It was found that the model correctly fits the data ($G^2 = 2,092.1634, p-value < 0.0001$). Based on the fitted model, Table 5.1 gives the estimated odds ratios for each effect, their associated 95% confidence intervals, as well as the p-values obtained when testing for the significance of each effect (Wald's test was used).

Table 5.1: Preliminary logistic regression results for Model 1

Factor	Odds ratio	95% C.I.	P-value
Year of collision			
2000	1.00		
2001	1.06	(0.82,1.36)	0.6564
2002	0.99	(0.77,1.28)	0.9474
2003	0.98	(0.76,1.27)	0.8722
2004	1.16	(0.90,1.49)	0.2559
2005	1.10	(0.85,1.42)	0.4632
Month of collision			
January	1.00		
February	0.85	(0.55,1.32)	0.4692
March	1.02	(0.66,1.56)	0.9436
April	1.15	(0.75,1.77)	0.5269
May	1.61	(1.08,2.41)	0.0195
June	1.38	(0.92,2.07)	0.1153
July	1.34	(0.90,2.00)	0.1492
August	1.50	(1.01,2.22)	0.0422
September	1.56	(1.05,2.32)	0.0285

Table 5.1: (continued)

Factor	Odds ratio	95% C.I.	P-value
October	1.08	(0.72,1.62)	0.7079
November	1.55	(1.06,2.27)	0.0244
December	1.29	(0.89,1.88)	0.1842
Day of the week			
Monday to Friday	1.00		
Saturday or Sunday	0.89	(0.75,1.05)	0.1708
Hour of collision			
0:00-2:59	1.00		
3:00-5:59	0.54	(0.20,1.48)	0.2272
6:00-8:59	0.66	(0.34,1.27)	0.2134
9:00-11:59	0.65	(0.35,1.20)	0.1670
12:00-14:59	0.60	(0.33,1.12)	0.1091
15:00-17:59	0.66	(0.36,1.21)	0.1826
18:00-20:59	0.53	(0.29,0.97)	0.0385
21:00-23:59	0.44	(0.23,0.84)	0.0132
Number of vehicles involved	1.12	(1.07,1.17)	<0.0001
Collision configuration			
Vehicle hit an object	1.00		
Other single vehicle configuration	1.01	(0.49,2.08)	0.9854
Rear-end collision	0.53	(0.24,1.19)	0.1253
Two vehicle, same direction of travel configuration other than rear-end collision	0.97	(0.44,2.15)	0.9385
Head-on collision	4.82	(2.27,10.25)	<0.0001
Right angle collision	1.52	(0.70,3.29)	0.2935

Table 5.1: (continued)

Factor	Odds ratio	95% C.I.	P-value
Two vehicle, different direction configuration other than head-on or right angle collision	1.24	(0.57,2.70)	0.5874
Roadway configuration			
Non-intersection	1.00		
At an intersection of at least two public roadways	0.63	(0.47,0.85)	0.0024
Intersection with a parking lot entrance/exit, private driveway or laneway	0.76	(0.56,1.05)	0.0958
Other	0.34	(0.13,0.89)	0.0277
Weather condition			
Clear and sunny, overcast or cloudy but no precipitation	1.00		
Raining	0.81	(0.56,1.17)	0.2564
Snowing, freezing rain, sleet, hail	0.77	(0.51,1.15)	0.2002
Visibility limitation	1.09	(0.65,1.82)	0.7501
Strong wind	1.11	(0.44,2.77)	0.8293
Light condition			
Daylight	1.00		
Dawn or dusk	1.27	(0.84,1.91)	0.2584
Darkness	1.20	(0.88,1.65)	0.2541
Road classification I			
Urban	1.00		
Rural	2.98	(2.36,3.76)	<0.0001
Road surface			

Table 5.1: (continued)

Factor	Odds ratio	95% C.I.	P-value
Dry or normal	1.00		
Wet	1.15	(0.86,1.54)	0.3632
Snow/Slush/Wet snow/Icy	0.76	(0.53,1.11)	0.1527
Other	1.26	(0.63,2.52)	0.5161
Road alignment			
Straight	1.00		
Curved	1.45	(1.18,1.79)	0.0004
Traffic control			
Traffic lights	1.00		
Sign	1.82	(1.40,2.38)	<0.0001
Other	1.88	(0.95,3.70)	0.0699
No control present	1.35	(0.98,1.87)	0.0689
Posted speed limit			
50 km/h or less	1.00		
60 to 90 km/h	2.25	(1.75,2.91)	<0.0001
100 km/h or more	2.99	(2.12,4.21)	<0.0001
Vehicle type			
Passenger car	1.00		
Passenger van or SUV	1.07	(0.85,1.36)	0.5647
Light truck	0.89	(0.69,1.14)	0.3507
Heavy truck	2.06	(1.13,3.74)	0.0180
Bus	2.42	(0.85,6.91)	0.0994
Motorcycle	4.40	(1.86,10.39)	0.0007
Other	1.55	(0.61,3.90)	0.3569
Vehicle manoeuvre			

Table 5.1: (continued)

Factor	Odds ratio	95% C.I.	P-value
Going straight ahead	1.00		
Turning left	1.14	(0.85,1.55)	0.3839
Turning right	0.62	(0.32,1.22)	0.1682
Making U-turn	0.84	(0.32,2.21)	0.7296
Changing lanes	0.73	(0.36,1.48)	0.3854
Reversing	1.49	(0.58,3.83)	0.4114
Overtaking or passing	0.56	(0.22,1.41)	0.2183
Slowing or stopping in traffic	0.46	(0.29,0.74)	0.0012
Other	0.55	(0.25,1.23)	0.1464
First impact location			
Front	1.00		
Roof	0.90	(0.50,1.59)	0.7059
Rear	0.55	(0.32,0.97)	0.0391
Left side	1.23	(1.03,1.47)	0.0252
Right side	1.04	(0.86,1.27)	0.6680
Other	0.85	(0.47,1.55)	0.5947
Vehicle event 1			
Non-collision event	1.00		
Hit moving object	1.19	(0.85,1.69)	0.3140
Hit non-moving object	1.01	(0.68,1.47)	0.9771
Driver sex			
Female	1.00		
Male	1.06	(0.90,1.25)	0.4804
Driver age	1.05	(1.03,1.06)	<0.0001
Vehicle age	1.01	(0.99,1.02)	0.1074

Table 5.1: (continued)

Factor	Odds ratio	95% C.I.	P-value
Number of restrained victims in the collision involving the studied driver	2.02	(1.80,2.27)	<0.0001
Number of unrestrained victims in the collision involving the studied driver	13.56	(10.32,17.81)	<0.0001
Number of victims who were not occupants of a vehicle in the collision involving the studied driver	5.56	(4.08,7.58)	<0.0001
Driver was fatigued or fell asleep	0.80	(0.50,1.29)	0.3669
Driver was under the effect of alcohol	2.58	(1.72,3.86)	<0.0001
Driver was under the effect of drugs	11.21	(4.88,25.72)	<0.0001
Other driver condition	1.63	(1.19,2.22)	0.0021
Following too closely	0.30	(0.13,0.66)	0.0030
Distraction, inattentiveness	0.80	(0.64,0.99)	0.0360
Driving too fast for conditions	1.14	(0.79,1.65)	0.4811
Improper turning or passing	0.79	(0.56,1.12)	0.1834
Failing to yield right-of-way	1.01	(0.79,1.29)	0.9617
Disobeying traffic control device or traffic officer	1.72	(1.28,2.31)	0.0003
Driving on wrong side of the road	0.53	(0.23,1.22)	0.1374
Lost control	1.15	(0.87,1.52)	0.3390
Other driver action	1.24	(0.93,1.65)	0.1434
Vehicle condition	0.84	(0.38,1.84)	0.6568
Environmental condition	1.27	(0.87,1.85)	0.2140
Other contributing factor	0.54	(0.29,1.02)	0.0563

However, this model contains many variables which do not have significant effects at the level $\alpha = 5\%$. Table 5.1 shows that the following variables have some effects that are significant at the level $\alpha = 5\%$ when modeling the risk of fatal collisions compared to non-fatal injury collisions: Month of collision, Hour of collision, Number of vehicles involved, Collision configuration, Roadway configuration, Road classification I, Road alignment, Traffic control, Posted speed limit, Vehicle type, Vehicle manoeuvre, First impact location, Driver age, Number of unrestrained victims in the collision involving the studied driver, Number of restrained victims in the collision involving the studied driver, Number of victims who were not occupants of a vehicle in the collision involving the studied driver, Driver was under the effect of alcohol, Driver was under the effect of drugs, Other driver condition, Following too closely, Distraction, inattentiveness, Disobeying traffic control device or traffic officer and Other driver condition. Forward selection (with a significance level of $\alpha = 5\%$) was then used to select the most significant variables among these 23 variables in a stepwise fashion. Table 5.2 shows the variable added to the model using forward selection according to the order in which they were added along with the p-value for each variable. Note that the variables added first are the most important.

Table 5.2: Variables selected in Model 1 using forward selection

Variable	df	χ^2	P-value
Number of unrestrained victims in the collision involving the studied driver	1	1,086.63	<0.0001
Collision configuration	6	915.98	<0.0001
Road classification I	1	597.57	<0.0001
Number of restrained victims in the collision involving the studied driver	1	94.75	<0.0001

Table 5.2: (continued)

Variable	df	χ^2	P-value
Number of victims who were not occupants of a vehicle in the collision involving the studied driver	1	162.51	<0.0001
Driver age	1	58.03	<0.0001
Driver was under the effect of drugs	1	42.17	<0.0001
Posted speed limit	2	41.48	<0.0001
Driver was under the effect of alcohol	1	27.36	<0.0001
Number of vehicles involved	1	23.36	<0.0001
Other driver condition	1	16.68	<0.0001
Vehicle type	6	29.60	<0.0001
Traffic control	3	20.83	0.0001
Disobeying traffic control device or traffic officer	1	13.78	0.0002
Road alignment	1	11.63	0.0006
Roadway configuration	3	15.66	0.0013
Month of collision	11	28.36	0.0029
Vehicle manoeuvre	8	23.19	0.0031
Following too closely	1	8.91	0.0028
First impact location	5	12.52	0.0284

Table 5.3 gives the estimated odds ratios for each effect of the variables entered in Model 1, with their associated 95% confidence intervals, as well as the p-values obtained when testing for the significance of each effect (Wald's test was used).

Table 5.3: Logistic regression results for Model 1 using forward selection

Factor	Odds ratio	95% C.I.	P-value
Month of collision			
January	1.00		
February	0.85	(0.55,1.33)	0.4781
March	1.03	(0.68,1.58)	0.8826
April	1.19	(0.79,1.80)	0.4089
May	1.72	(1.18,2.51)	0.0048
June	1.48	(1.01,2.16)	0.0423
July	1.42	(0.98,2.06)	0.0639
August	1.61	(1.12,2.32)	0.0109
September	1.67	(1.15,2.42)	0.0077
October	1.17	(0.80,1.73)	0.4158
November	1.62	(1.12,2.35)	0.0110
December	1.27	(0.88,1.85)	0.2087
Number of vehicles involved	1.11	(1.06,1.16)	<0.0001
Collision configuration			
Vehicle hit an object	1.00		
Other single vehicle configuration	0.94	(0.47,1.87)	0.8501
Rear-end collision	0.57	(0.27,1.23)	0.1504
Two vehicle, same direction of travel configuration other than rear-end collision	1.00	(0.47,2.11)	0.9933
Head-on collision	4.69	(2.32,9.49)	<0.0001
Right angle collision	1.52	(0.74,3.14)	0.2553

Table 5.3: (continued)

Factor	Odds ratio	95% C.I.	P-value
Two vehicle, different direction configuration other than head-on or right angle collision	1.29	(0.62,2.67)	0.4996
Roadway configuration			
Non-intersection	1.00		
At an intersection of at least two public roadways	0.63	(0.47,0.85)	0.0025
Intersection with a parking lot entrance/exit, private driveway or laneway	0.78	(0.57,1.07)	0.1292
Other	0.33	(0.13,0.85)	0.0213
Road classification I			
Urban	1.00		
Rural	2.91	(2.31,3.67)	<0.0001
Road alignment			
Straight	1.00		
Curved	1.40	(1.14,1.72)	0.0012
Traffic control			
Traffic lights	1.00		
Sign	1.81	(1.39,2.35)	<0.0001
Other	1.81	(0.92,3.56)	0.0849
No control present	1.30	(0.94,1.80)	0.1079
Posted speed limit			
50 km/h or less	1.00		
60 to 90 km/h	2.28	(1.77,2.94)	<0.0001
100 km/h or more	2.97	(2.12,4.16)	<0.0001

Table 5.3: (continued)

Factor	Odds ratio	95% C.I.	P-value
Vehicle type			
Passenger car	1.00		
Passenger van or SUV	1.08	(0.85,1.36)	0.5279
Light truck	0.91	(0.71,1.15)	0.4197
Heavy truck	2.21	(1.23,3.99)	0.0083
Bus	2.37	(0.83,6.78)	0.1084
Motorcycle	4.65	(1.98,10.88)	0.0004
Other	1.62	(0.65,4.06)	0.3018
Vehicle manoeuvre			
Going straight ahead	1.00		
Turning left	1.03	(0.79,1.35)	0.8154
Turning right	0.57	(0.29,1.11)	0.0981
Making U-turn	0.72	(0.28,1.85)	0.4991
Changing lanes	0.64	(0.33,1.27)	0.2002
Reversing	1.47	(0.58,3.76)	0.4179
Overtaking or passing	0.50	(0.20,1.24)	0.1332
Slowing or stopping in traffic	0.47	(0.29,0.74)	0.0011
Other	0.51	(0.23,1.15)	0.1033
First impact location			
Front	1.00		
Roof	0.88	(0.50,1.54)	0.6423
Rear	0.56	(0.32,0.97)	0.0384
Left side	1.23	(1.03,1.47)	0.0216
Right side	1.03	(0.85,1.24)	0.7882
Other	0.84	(0.47,1.50)	0.5489

Table 5.3: (continued)

Factor	Odds ratio	95% C.I.	P-value
Driver age	1.05	(1.03,1.06)	<0.0001
Number of restrained victims in the collision involving the studied driver	1.97	(1.75,2.20)	<0.0001
Number of unrestrained victims in the collision involving the studied driver	12.84	(9.83,16.78)	<0.0001
Number of victims who were not occupants of a vehicle in the collision involving the studied driver	6.00	(4.56,7.90)	<0.0001
Driver was under the effect of alcohol	2.80	(1.90,4.13)	<0.0001
Driver was under the effect of drugs	11.28	(4.98,25.58)	<0.0001
Other driver condition	1.79	(1.32,2.43)	0.0002
Following too closely	0.28	(0.13,0.63)	0.0019
Disobeying traffic control device or traffic officer	1.66	(1.26,2.19)	0.0004

Model 1 shows that many factors significantly increase the risk of a fatal collision. First, it shows that an increase of 1 in the number of unrestrained victims in the collision involving the studied driver increases the risk of a fatal collision by a factor of 12.84 times (95% C.I.: 9.83-16.78), but that the risk increases by only 1.97 times (C.I.: 1.75-2.20) when the number of restrained victims in the collision involving the studied driver increases by 1 unit. Thus, collisions with unrestrained victims are much more likely to be fatal than collisions with restrained victims. An increase of 1 in the number of victims who were not occupants of one of the vehicles in the collision involving the studied driver increases the risk of a fatal collision by 6 times (C.I.: 4.56-7.90).

Collisions where the driver was under the effect of drugs are 11.28 times (C.I.: 4.98-25.58) more likely to be fatal collisions than collisions where the driver was not under the effect of drugs. Moreover, collisions where the driver was under the effect of alcohol are 2.80 times (C.I.: 1.90-4.13) more likely to be fatal collisions than collisions where the driver was not under the effect of alcohol.

Model 1 also shows that head-on collisions are 4.69 times (C.I.: 2.32-9.49) more likely to be fatal collisions than collisions where the vehicle hit an object. Collisions where the vehicle of the studied driver is a motorcycle are 4.65 times (C.I.: 1.98-10.88) more likely to be fatal collisions than collisions where the vehicle of the studied driver is a passenger car. Collisions where the vehicle of the studied driver is a heavy truck are 2.21 times (C.I.: 1.23-3.99) more likely to be fatal collisions than collisions where the vehicle of the studied driver is a passenger car.

Collisions where the posted speed limit is between 60 and 90 km/h are 2.28 times (C.I.: 1.77-2.94) more likely to be fatal collisions than collisions where the posted speed limit is below 60 km/h and collisions where the posted speed limit is at least 100 km/h are 2.97 times (C.I.: 2.12-4.16) more likely to be fatal collisions. Collisions on rural roads are 2.91 times (C.I.: 2.31-3.67) more likely to be fatal collisions than collisions on urban roads.

Collisions where signs are used to control traffic at the site of the collision are 81% (C.I.: 1.39-2.35) more likely to be fatal collisions than collisions where traffic lights are used. Collisions where the driver had another condition than those mentioned are 79% (C.I.: 1.32-2.43) more likely to be fatal collisions than collisions where the driver did not. Collisions where the driver disobeyed traffic control device or traffic officer are 66% (C.I.: 1.26-2.19) more likely to be fatal collisions than collisions where the driver did not. Collisions occurring on roads curved at the collision site are also 40% (C.I.: 1.14-1.72) more likely to be fatal collisions than collisions where the road is going straight at the collision site. Collisions where the first impact is on the left side of the vehicle are 23% (C.I.: 1.03-1.47) more likely to be fatal collisions than

collisions where the first impact is in the front of the vehicle.

The model also shows that an increase of 1 vehicle in the number of vehicles involved in the collisions increases the risk of a collision to be fatal by 11% (C.I.: 1.06-1.16). An increase of 1 year in the age of the older driver increases the risk of fatal collision by 5% (C.I.: 1.03-1.06). Finally, collisions occurring during the months of May, June, August, September and November are respectively 72% (C.I.: 1.18-2.51), 48% (C.I.: 1.01-2.16), 61% (C.I.: 1.12-2.32), 67% (C.I.: 1.15-2.42) and 62% (C.I.: 1.12-2.35) more likely to be fatal collisions than collisions occurring in January.

5.2 Modeling the risk of fatal injuries to the driver

The risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older was also modeled. As only the cases without missing values are used, 24,723 drivers were selected for analysis. 478 fatally injured drivers and 24,245 non-fatally injured drivers were included in the analysis. Again, a preliminary model using all the predictor variables has been fitted. The model obtained is called Model 2. The fitted model was tested for goodness-of-fit using the likelihood-ratio test. It was found that the model fits correctly the data ($G^2 = 1,111.9920, p - value < 0.0001$). Based on the fitted model, Table 5.4 gives the estimated odds ratios for each effect, their associated 95% confidence intervals, as well as the p-values obtained when testing for the significance of each effect (Wald's test was used).

Table 5.4: Preliminary logistic regression results for Model 2

Factor	Odds ratio	95% C.I.	P-value
Year of collision			
2000	1.00		
2001	1.18	(0.86,1.63)	0.3068
2002	0.98	(0.71,1.36)	0.9186

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
2003	0.87	(0.62,1.22)	0.4130
2004	1.16	(0.84,1.60)	0.3764
2005	0.46	(0.31,0.67)	<0.0001
Month of collision			
January	1.00		
February	1.02	(0.59,1.76)	0.9424
March	1.10	(0.64,1.90)	0.7342
April	1.02	(0.58,1.80)	0.9335
May	1.41	(0.84,2.37)	0.1985
June	1.09	(0.64,1.87)	0.7501
July	1.20	(0.71,2.02)	0.4916
August	1.38	(0.83,2.29)	0.2145
September	1.08	(0.62,1.85)	0.7954
October	1.00	(0.59,1.70)	0.9896
November	1.45	(0.88,2.39)	0.2006
December	1.36	(0.84,2.20)	0.1484
Day of the week			
Monday to Friday	1.00		
Saturday or Sunday	0.92	(0.73,1.16)	0.4733
Hour of collision			
0:00-2:59	1.00		
3:00-5:59	0.85	(0.28,2.60)	0.7799
6:00-8:59	0.45	(0.19,2.05)	0.0635
9:00-11:59	0.46	(0.21,1.02)	0.0557
12:00-14:59	0.36	(0.16,0.80)	0.0117

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
15:00-17:59	0.48	(0.22,1.04)	0.0627
18:00-20:59	0.42	(0.19,0.90)	0.0261
21:00-23:59	0.53	(0.24,1.21)	0.1308
Number of vehicles involved	1.10	(0.98,1.24)	0.0999
Collision configuration			
Vehicle hit an object	1.00		
Other single vehicle configuration	1.28	(0.53,3.14)	0.5833
Rear-end collision	0.89	(0.31,2.56)	0.8230
Two vehicle, same direction of travel configuration other than rear-end collision	0.99	(0.35,2.79)	0.9799
Head-on collision	5.56	(2.14,14.44)	0.0004
Right angle collision	1.63	(0.61,4.36)	0.3298
Two vehicle, different direction configuration other than head-on or right angle collision	1.52	(0.56,4.12)	0.4088
Roadway configuration			
Non-intersection	1.00		
At an intersection of at least two public roadways	0.66	(0.44,1.00)	0.0477
Intersection with a parking lot entrance/exit, private driveway or laneway	0.85	(0.56,1.30)	0.4515
Other	0.37	(0.12,1.11)	0.0749
Weather condition			
Clear and sunny, overcast or cloudy but no precipitation	1.00		

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
Raining	0.79	(0.50,1.25)	0.3167
Snowing, freezing rain, sleet, hail	0.78	(0.47,1.30)	0.3401
Visibility limitation	0.80	(0.39,1.63)	0.5393
Strong wind	0.58	(0.13,2.63)	0.4803
Light condition			
Daylight	1.00		
Dawn or dusk	0.95	(0.54,1.68)	0.8622
Darkness	0.86	(0.56,1.33)	0.5025
Road classification I			
Urban	1.00		
Rural	2.36	(1.73,3.22)	<0.0001
Road surface			
Dry or normal	1.00		
Wet	1.33	(0.92,1.92)	0.1273
Snow/Slush/Wet snow/Icy	0.85	(0.53,1.35)	0.4844
Other	1.54	(0.65,3.67)	0.3320
Road alignment			
Straight	1.00		
Curved	1.32	(1.00,1.73)	0.0483
Traffic control			
Traffic lights	1.00		
Sign	2.33	(1.62,3.35)	<0.0001
Other	1.71	(0.56,5.20)	0.3452
No control present	1.43	(0.90,2.25)	0.1301
Posted speed limit			

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
50 km/h or less	1.00		
60 to 90 km/h	2.36	(1.67,3.34)	<0.0001
100 km/h or more	3.61	(2.28,5.71)	<0.0001
Vehicle type			
Passenger car	1.00		
Passenger van or SUV	1.00	(0.72,1.41)	0.9863
Light truck	0.85	(0.61,1.18)	0.3254
Heavy truck	1.11	(0.36,3.38)	0.8583
Bus	1.59	(0.20,12.51)	0.6583
Motorcycle	0.17	(0.00,5.12)	0.9659
Other	0.60	(0.12,2.99)	0.5279
Vehicle manoeuvre			
Going straight ahead	1.00		
Turning left	1.04	(0.68,1.59)	0.8464
Turning right	0.44	(0.16,1.25)	0.1244
Making U-turn	0.85	(0.24,3.00)	0.8060
Changing lanes	1.22	(0.50,2.99)	0.6092
Reversing	1.65	(0.45,6.07)	0.4517
Overtaking or passing	0.84	(0.26,2.78)	0.7801
Slowing or stopping in traffic	0.47	(0.23,0.96)	0.0384
Other	0.82	(0.34,1.98)	0.6543
First impact location			
Front	1.00		
Roof	1.14	(0.60,2.15)	0.6965
Rear	0.37	(0.16,0.89)	0.0254

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
Left side	1.61	(1.27,2.03)	<0.0001
Right side	0.82	(0.62,1.09)	0.1648
Other	1.02	(0.48,2.17)	0.9580
Vehicle event 1			
Non-collision event	1.00		
Hit moving object	0.74	(0.48,1.15)	0.1840
Hit non-moving object	1.06	(0.69,1.65)	0.7816
Driver sex			
Female	1.00		
Male	1.09	(0.88,1.35)	0.4306
Driver age	1.05	(1.04,1.07)	<0.0001
Vehicle age	1.01	(0.99,1.03)	0.1416
Safety device used			
Restrained	1.00		
Unrestrained	10.70	(8.16,14.03)	<0.0001
Driver was fatigued or fell asleep	0.84	(0.48,1.47)	0.5367
Driver was under the effect of alcohol	1.92	(1.20,3.08)	0.0068
Driver was under the effect of drugs	3.78	(1.33,10.74)	0.0126
Other driver condition	1.77	(1.23,2.56)	0.0021
Following too closely	0.16	(0.02,1.18)	0.0714
Distraction, inattentiveness	0.48	(0.35,0.66)	<0.0001
Driving too fast for conditions	1.39	(0.87,2.21)	0.1685
Improper turning or passing	0.85	(0.51,1.40)	0.5196
Failing to yield right-of-way	1.77	(1.27,2.47)	0.0008

Table 5.4: (continued)

Factor	Odds ratio	95% C.I.	P-value
Disobeying traffic control device or traffic officer	3.02	(2.06,4.41)	<0.0001
Driving on wrong side of the road	1.31	(0.50,3.46)	0.5845
Lost control	1.27	(0.89,1.80)	0.1891
Other driver action	1.34	(0.91,1.96)	0.1366
Vehicle condition	0.39	(0.09,1.61)	0.1922
Environmental condition	1.14	(0.69,1.90)	0.6033
Other contributing factor	0.66	(0.30,1.43)	0.2876

Table 5.4 show that the following variables have some effects which are significant at the level $\alpha = 5\%$ when modeling the risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older: Year or collision, Hour of collision, Collision configuration, Roadway configuration, Road classification I, Road alignment, Traffic control, Posted speed limit, Vehicle manoeuvre, First impact location, Driver age, Safety device used, Driver was under the effect of alcohol, Driver was under the effect of drugs, Other driver condition, Distraction, inattentiveness, Failing to yield the right-of-way and Disobeying traffic control device or traffic officer. Forward selection (with a significance level of $\alpha = 5\%$) was then used to select the most significant variables among these 18 variables. Table 5.5 shows the variables added to the model using forward selection according to the order in which they were added along with the p-value for each variable. Note that the variables added first are the most important.

Table 5.5: Variables selected in Model 2 using forward selection

Variable	df	χ^2	P-value
Variable	DF	χ^2	P-value
Safety device used	1	450.38	<0.0001
Road classification I	1	395.82	<0.0001
Collision configuration	6	187.58	<0.0001
Driver age	1	54.77	<0.0001
Year of collision	5	43.72	<0.0001
First impact location	5	38.92	<0.0001
Posted speed limit	2	26.63	<0.0001
Distraction, inattentiveness	1	15.58	<0.0001
Disobeying traffic control device or traffic officer	1	19.81	<0.0001
Traffic control	3	27.78	0.0001
Other driver condition	1	13.05	0.0003
Driver was under the effect of alcohol	1	12.26	0.0005
Failing to yield right-of-way	1	11.19	0.0008
Driver was under the effect of drugs	1	8.92	0.0028
Road alignment	1	4.82	0.0282

Table 5.6 gives the estimated odds ratios for each effect of the variables entered in Model 2, with their associated 95% confidence intervals, as well as the p-values obtained when testing for the significance of each effect (Wald's test was used).

Table 5.6: Logistic regression results for Model 2 using forward selection

Factor	Odds ratio	95% C.I.	P-value
Year of collision			
2000	1.00		
2001	1.16	(0.85,1.60)	0.3448
2002	0.97	(0.70,1.34)	0.8486
2003	0.86	(0.62,1.20)	0.3730
2004	1.17	(0.85,1.61)	0.3325
2005	0.42	(0.29,0.61)	<0.0001
Collision configuration			
Vehicle hit an object	1.00		
Other single vehicle configuration	1.19	(0.51,2.79)	0.6946
Rear-end collision	0.46	(0.17,1.22)	0.1183
Two vehicle, same direction of travel configuration other than rear-end collision	0.72	(0.28,1.89)	0.5082
Head-on collision	4.13	(1.72,9.90)	0.0015
Right angle collision	1.06	(0.44,2.59)	0.8967
Two vehicle, different direction configuration other than head-on or right angle collision	0.96	(0.40,2.33)	0.9267
Road classification I			
Urban	1.00		
Rural	2.40	(1.77,3.25)	<0.0001
Road alignment			
Straight	1.00		
Curved	1.35	(1.03,1.76)	0.0286

Table 5.6: (continued)

Factor	Odds ratio	95% C.I.	P-value
Traffic control			
Traffic lights	1.00		
Sign	2.32	(1.62,3.32)	<0.0001
Other	1.66	(0.56,4.89)	0.3606
No control present	1.79	(1.24,2.59)	0.0018
Posted speed limit			
50 km/h or less	1.00		
60 to 90 km/h	2.36	(1.68,3.32)	<0.0001
100 km/h or more	3.67	(2.36,5.71)	<0.0001
First impact location			
Front	1.00		
Roof	1.16	(0.62,2.17)	0.6362
Rear	0.37	(0.16,0.85)	0.0182
Left side	1.61	(1.28,2.03)	<0.0001
Right side	0.82	(0.62,1.09)	0.1714
Other	1.07	(0.51,2.24)	0.8599
Driver age	1.06	(1.04,1.07)	<0.0001
Safety device used			
Restrained	1.00		
Unrestrained	9.91	(7.64,12.85)	<0.0001
Driver was under the effect of alcohol	2.18	(1.39,3.41)	0.0007
Driver was under the effect of drugs	4.28	(1.53,11.94)	0.0056
Other driver condition	1.97	(1.38,2.80)	0.0002
Distraction, inattentiveness	0.48	(0.35,0.65)	<0.0001
Failing to yield right-of-way	1.64	(1.23,2.19)	0.0008

Table 5.6: (continued)

Factor	Odds ratio	95% C.I.	P-value
Disobeying traffic control device or traffic officer	2.93	(2.02,4.23)	<0.0001

Model 2 shows that many factors significantly increase the risk of fatal injuries to the driver. First, unrestrained drivers are 9.91 times (95% C.I.: 7.64-12.85) more likely to be fatally injured than restrained drivers. Drivers under the effect of drugs are 4.28 times (C.I.: 1.53-11.94) more likely to be fatally injured than drivers not under the effect of drugs. Drivers under the effect of alcohol are 2.18 times (C.I.: 1.39-3.41) more likely to be fatally injured than drivers not under the effect of alcohol. Drivers involved in head-on collisions are 4.13 times (C.I.: 1.72-9.90) more likely to be fatally injured than drivers involved in collisions where the vehicle hit an object.

Drivers who disobeyed traffic control device or a traffic officer at the collision time are 2.93 times (C.I.: 2.02-4.23) more likely to be fatally injured than drivers who did not. Drivers involved in collisions where the posted speed limit at the collision site is between 60 and 90 km/h are 2.36 times (C.I.: 1.68-3.32) more likely to be fatally injured than drivers involved in collisions where the posted speed limit is below 60 km/h and drivers involved in collisions where the posted speed limit at the collision site is at least 100 km/h are 3.67 times (C.I.: 2.36-5.71) more likely to be fatally injured.

Drivers involved in collisions which occurred on rural roads are 2.40 times (C.I.: 1.77-3.25) more likely to be fatally injured than drivers involved in collisions on urban roads. Drivers involved in collisions where signs are used to control traffic at the site of the collision are 2.32 times (C.I.: 1.62-3.32) more likely to be fatally injured than drivers involved in collisions where traffic lights are used. Drivers involved in collisions where no traffic control device was present at the site of the collision are 79% (C.I.:

1.24-2.59) more likely to be fatally injured than drivers involved in collisions where traffic lights are present at the collision site.

Drivers who suffered another condition than those mentioned previously are 97% (C.I.: 1.38-2.80) more likely to be fatally injured than drivers involved in collisions where the driver did not. Drivers who failed to yield the right-of-way at the time of the collisions are 64% (C.I.: 1.23-2.19) more likely to be fatally injured than drivers who did not.

Drivers involved in collisions where the first impact was on the left side of the vehicle are 61% (C.I.: 1.28-2.03) more likely to be fatally injured than drivers involved in collisions where the first impact is in the front of the vehicle. Drivers involved in collisions which occurred on roads curved at the collision site are 35% (C.I.: 1.03-1.76) more likely to be fatally injured than drivers involved in collisions where the road was going straight at the collision site. Finally, an increase of 1 year in the age of the driver increases the risk of the driver being fatally injured by 6% (C.I.: 1.04-1.07).

Chapter 6

Discussion

6.1 Limitations of the study

This study has some limitations mainly due to the data available. Many variables could not be included in the model, as mentioned in section 1.3. Unfortunately, some variables not used in the model, such as Vehicle speed, would have been interesting to include.

Moreover, as some observations have missing values and logistic regression can't handle these observations, the number of observations used in the logistic regression models is lower than the number of observations used in the classification trees. This could lead to some differences between the classification trees and the logistic regression models.

6.2 Comparison between classification trees and logistic regression

6.2.1 Risk of fatal collisions

The classification tree and the logistic regression model used to model the risk of fatal collisions compared to non-fatal injury collisions among collisions involving drivers

aged 65 years and older have both identified the following variables as having significant effects on predicting severity of the collisions:

- Road classification I
- Collision configuration
- Posted speed limit
- Vehicle manoeuvre
- Vehicle type
- Driver age
- Number of vehicles involved
- Number of unrestrained victims in the collision involving the studied driver
- Number of restrained victims in the collision involving the studied driver
- Road alignment.

The variables Year of collision, Hour of collision and Vehicle age are used in the best pruned tree but are not significant in the logistic regression model. On the other hand, some variables not used in the tree are also significant in the logistic regression model. These variables are: First impact location, Number of victims who were not occupants of a vehicle in the collision involving the studied driver, Driver was under the effect of drugs, Driver was under the effect of alcohol, Other driver condition, Traffic control, Disobeying traffic control device or traffic officer, Roadway configuration, Month of collision, First impact location and Following too closely.

Note that some of the variables not used in the tree but used in the logistic regression model, such as Traffic control, Disobeying traffic control device or traffic officer, Roadway configuration, Month of collision and Following too closely, were among the last variables added to the logistic regression model using forward selection. The absence in the tree of some variables significant in the logistic regression model can be also be due to the differences between the two methods. However, in classification trees, the effects of some variables can be masked by other variables. As

an example, suppose the first split uses some given variable as the splitting variable. This variable is the variable on which the split reducing the most the impurity at this point is based. However, some other splits, based on different variables, could also reduce significantly the impurity of the tree. The effects of the variables used in these splits are thus possibly masked as they are not necessarily used further into the tree.

Fortunately, S-Plus also provides, for each split in the tree, the splits producing the greatest impurity reductions after the best split, giving an indication of which variables could have effects masked by the variables used in the tree. By looking at these splits, we find that most of the significant variables having significant effects in the logistic regression model are among the best 5 splits for some split present in the tree. As an example, the variable Traffic control is the fifth best splitting variable for node 1 and would also reduce significantly the impurity if it was used.

Note also that the first 4 variables added in the logistic regression model, Number of unrestrained victims in the collision involving the studied driver, Number of restrained victims in the collision involving the studied driver, Collision configuration and Road classification 1, appear in the first 3 levels of the the classification tree, although in a different order.

6.2.2 Risk of fatal injuries to the driver

The classification tree and the logistic regression model used to model the risk of fatal injuries compared to non-fatal injuries for drivers aged 65 years and older have both identified the following variables as having significant effects on predicting the severity of the injuries to the drivers:

- Road classification I
- Collision configuration
- Safety device used
- Driver age
- Other driver condition

- Year of collision
- Traffic control.

The variables Driver sex, Weather condition, Improper turning, Month of collision, Hour of collision, Vehicle manoeuvre, Vehicle age and Vehicle type are used in the best pruned tree but are not significant in the logistic regression model. On the other hand, some variables not used in the tree are also significant in the logistic regression model. These variables are: Posted speed limit, Distraction, inattentiveness, Disobeying traffic control device or traffic officer, Driver was under the effect of alcohol, Failing to yield right-of-way, Driver was under the effect of drugs, Road alignment and First impact location.

Again, some of the variables not used in the tree but used in the logistic regression model, such as Driver was under the effect of alcohol, Failing to yield right-of-way, Driver was under the effect of drugs or Road alignment, were the last variables added to the logistic regression model using forward selection. The absence in the tree of some variables significant in the logistic regression model can be also be due to the differences between the two methods. Again, some variables seem masked by others in the classification tree. Based on the S-plus output, we find that most of the significant variables in the logistic regression model are among the best 5 splits for some splits present in the tree. As an example, the variable Posted speed limit is the second best splitting variable for node 1 and would reduce significantly the impurity if it was used.

Note that the first 4 variables added in the logistic regression model, Safety device used, Road classification 1, Collision configuration and Driver age, appear in the first 3 levels of the classification tree, although in a different order.

These results, particularly those of the model for the risk of fatal collisions, show that even though classification trees and logistic regression models are different in many ways, their results contain many similarities. However, this can't necessarily be generalized for all kind of data.

Note also that a logistic regression model including all the interactions between variables would probably give results closer to the classification tree. However, when there is a large number of predictor variables in the model, it is difficult to do so.

6.3 Advantages and disadvantages of classification trees and logistic regression

Both classification trees and logistic regression have their own advantages and disadvantages. We will take a closer look at these.

One of the greatest advantages of classification trees is that they are easier to interpret than logistic regression models, especially for non-statisticians. A second advantage is that classification trees use all the available observations as they have a mechanism to deal with missing values, whereas logistic regression does not use the observations with missing values. A third advantage of classification trees is that they show clearly the interactions between the variables as opposed to logistic regression where higher-order terms need to be added to take into account the interactions. Adding such terms could increase the size of the model rapidly, especially in problems with many predictor variables.

On the other hand, an important advantage of logistic regression is that it provides a statistical interpretation of the relationship between the response variable and the predictor variables by providing, among other things, odds ratios for each effect in the model and p-values for testing the significance of each effect.

6.4 Comparison between the two models

The following factors were shown to increase significantly the risk of fatal collisions among collisions involving drivers aged 65 years and older by both the classification tree and the logistic regression model:

- Increasing number of unrestrained victims in the collision involving the studied driver
- Collision occurred on a rural road
- Posted speed limit is at least 60 km/h
- Head-on collisions
- Vehicle of the studied driver is a heavy truck or a motorcycle
- Collision occurred on a road curved at the collision site
- Increasing number of vehicles involved in the collision
- Increasing driver age.

Moreover, the following factors were identified as factors increasing significantly the risk of fatal injuries for drivers aged 65 years and older by both the classification tree and the logistic regression model:

- Non-use of restraint
- Collision occurred on a rural road
- Head-on collision
- Increasing driver age
- Signs are used to control traffic at the site of the collision
- No traffic control device is present at the site of the collision
- Driver suffered a condition other than being under the effect of alcohol or drugs or falling asleep while driving.

These two models have some important differences as one is modeling the severity of the collisions and the other is modeling the severity of the injuries to the driver. Higher posted speed limits, an increasing number of vehicles involved in the collision, the fact that the collision occurred on a road curved at the site of the collision and the fact that the vehicle of the studied driver is an heavy truck or a motorcycle are increasing significantly the risk of fatal collisions but not the risk of fatal injuries to

the driver. On the other hand, the fact that signs are used to control the traffic at the collision site or that no control device is present or the fact that a driver suffered a condition other than being under the effect of alcohol or drugs or falling asleep while driving are increasing significantly the risk of fatal injuries to the driver but not the risk of fatal collisions.

These differences can be due to the fact that the number of collisions used in the first model is much higher than the number of drivers included in the second model, thus producing more significant results for the first model. This can also be due to more fundamental differences in the two predictor variables. As an example, drivers of heavy trucks are not necessarily killed in a collision, but they are more likely to kill the occupants of another vehicle or the non-occupants than if they drove a passenger car.

Moreover, some odds ratios in the logistic regression model for the risk of fatal injuries to the driver are lower than in the model for the risk of fatal collisions. As an example, collisions where the driver is under the effect of drugs are 11.28 times more likely to be fatal, but drivers under the effect of drugs involved in collisions are 4.28 times more likely to be fatally injured. Therefore, the model using only drivers doesn't provide a picture of the problem that take into account the victims as the first model.

To try to reduce the number of fatalities among the collisions involving drivers aged 65 years and older, the following factors are the most important to look at, according to their odds ratios in the model for the risk of fatal collisions:

- Number of unrestrained victims
- Driver was under the effect of drugs
- Head-on collisions
- Vehicle of the studied driver is a motorcycle
- Posted speed limit was at least 100 km/h
- Collision occurred on a rural road

- Driver was under the effect of alcohol
- Posted speed limit was between 60 and 90 km/h
- Vehicle of the studied driver is a truck.

6.5 Comparison with previous studies

We will first compare the results of the logistic regression model modeling the risk of fatal collisions compared to non-fatal injury collisions to those of the study by Zhang, Lindsay, Clarke, Robbins and Mao (1999)[2]. Note that even though both studies show some similarities, there are also many differences. The group of collisions selected for the analysis is not exactly the same. Moreover, their model was comparing fatal collisions to minimal-injury collisions and includes fewer variables.

The following factors were found to be significant in both their study and in our model: driver age, non-use of seat belts, intersection without traffic controls, roads with higher speed limits and head-on collisions. However, their study also showed that sex, failing to yield right-of-way/disobeying traffic signs, snowy weather, two-vehicle turning collisions, overtaking and changing lanes had a significant effect on the severity of the collisions, whereas our results did not.

Now, we will compare the results of the logistic regression model modeling the risk of fatal injuries to the driver to those of other studies.

Dissanayake and Lu (2001)[3] found that the following variables affected the severity of the injuries to drivers aged 65 years and older: travel speed, restraint device usage, point of impact, use of alcohol and drugs, personal condition, gender, whether the driver is at fault, urban/rural nature of the road and grade/curve existence at the collision location. All these factors were shown to be significant in our model too with the exception of point of impact, gender and whether the driver is at fault. NCDB does not include the latter factor.

Moreover, Hill and Boyle (2005)[4], Bédard, Guyatt, Stones and Hirdes (2001)[5], Mercier, Shelley, Rimkus and Mercier (1997)[6] showed that the risk of serious or fatal

injuries to drivers significantly increases with age, similar to the results we obtained.

To our knowledge, no study involving classification trees has been done before on this subject.

Chapter 7

Conclusion

The primary purpose of this thesis was to model the relationship between the severity of the collisions and their characteristics. Based on the analysis of the data, the following factors were shown to increase significantly the risk of fatal collisions among collisions involving drivers aged 65 years and older by both the classification tree and the logistic regression model:

- Increasing number of unrestrained victims in the collision involving the studied driver
- Collision occurred on a rural road
- Posted speed limit is at least 60 km/h
- Head-on collisions
- Vehicle of the studied driver is a heavy truck or a motorcycle
- Collision occurred on a road curved at the collision site
- Increasing number of vehicles involved in the collision
- Increasing driver age.

Moreover, the following factors were identified as factors increasing significantly the risk of fatal injuries for drivers aged 65 years and older by both the classification tree and the logistic regression model:

- Non-use of restraint
- Collision occurred on a rural road
- Head-on collision
- Increasing driver age
- Signs are used to control traffic at the site of the collision
- No traffic control device is present at the site of the collision
- Driver suffered another condition than being under the effect of alcohol or drugs or falling asleep while driving.

Other factors have been identified to be important by either one of the techniques but not by both.

Another goal was to compare classification trees to logistic regression. Depending on the use one wants to make of the model obtained, both methods have their advantages. In this study, we have validated the classification tree with the more rigorous regression analysis. Consequently, the non-statistician can use the visually appealing trees with confidence.

In broader applications, both methods could be compared using cross-validation to find the distribution of the error for both techniques.

Bibliography

[1] Lécuyer, J., Chouinard, A., Hurley, R., Les effets des changements démographiques sur le nombre de collisions, *Canadian Multidisciplinary Road Safety Conference*, Fredericton, 2005

[2] Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y., Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario, *Accident Analysis and Prevention* 32, 117-125 (2000).

[3] Dissanayake, S., Lu, J.J, Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes, *Accident Analysis and Prevention* 34, 609-618 (2002).

[4] Hill, J.D., Boyle, L.N. Analyzing Severe Injury Risk for Crashes Nationally and Within Iowa, in: *Proceedings of the 2005 Mid-continent Transportation Research Symposium* (2005).

[5] Bédard, M., Guyatt, G.H., Stones, M.J, Hirdes, J.P., The independent contribution of driver, crash, and vehicle characteristics to driver fatalities, *Accident Analysis and Prevention* 34, 717-727 (2002).

[6] Mercier, C.R., Shelley II, M.C., Rimkus, J.B., Mercier, J.M., Age and Gender as Predictors of Injury Severity in Head-on Highway Vehicular Collisions, *Transportation Research Record* 1581, TRB, 37-46 (2001).

[7] Transport Canada, *National collision database (NCDB)*, 2000-2005.

[8] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, CRC Press, 1984.

[9] Therneau, T.M., Atkinson, E.J., *An introduction to recursive partionning using the RPART routines*, Technical report 61, Mayo Clinic, Section of statistics, 1997.

[10] Marsden, J.E., Tromba, A.J., *Vector calculus*, Freeman, 1996.

[11] Agresti, A., *Catgorical Data Analysis*, Wiley-Interscience, 2002.

[12] McCullagh, P., Nelder, J.A., *Generalized Linear Models*, Chapman and Hall, 1989.

[13] Neter, J., Kutner, M.H., Nachsteim, C.J., Wasserman, W., *Applied Linear Statistical Models*, McGraw-Hill, 1996.