Université de Montréal

BART applied to insurance

par

Catherine Paradis-Therrien

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

octobre 2007

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

BART applied to insurance

présenté par

Catherine Paradis-Therrien

a été évalué par un jury composé des personnes suivantes :

Charles Dugas
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Yves Lepage
(membres du jury)

Mémoire accepté le:
3 octobre 2007

SOMMAIRE

Ce mémoire porte sur deux méthodes de forage de données (data mining) appliquées au domaine de l'assurance. Afin de mieux comprendre les données, deux méthodes d'analyse en grappes soient les analyses hiérarchiques et non hiérarchiques, sont utilisées. Ensuite, des arbres de décision sont développés en utilisant le ratio de vente comme variable dépendante. Le but de ces modèles est de prédire les acheteurs de produits d'assurance les plus probables. Pour ce faire, les algorithmes de construction selon l'approche classique et l'approche bayésienne sont confrontés. Ainsi, l'approche classique utilise un algorithme de construction qui minimise une fonction d'impureté lors de chaque séparation. L'approche bayésienne quant à elle utilise plusieurs distributions *a priori* telles que celles des variables et du nombre de noeuds terminaux. L'objectif dans ce cas est de trouver l'arbre ayant la probabilité *a posteriori* la plus grande possible. Une fois que les arbres ont été construits selon les deux approches, les résultats sont comparés afin de déterminer quelle est celle qui donne les meilleurs arbres.

SUMMARY

This Masters thesis presents two data mining techniques applied in the insurance business. First, hierarchical and partitional clustering are used to have a better knowledge of the population under consideration. Then, in order to predict the most potential buyers, we consider the decision tree models with the closing ratio as the target variable. These trees are developed using the classical and the Bayesian statistical approaches. The classical method algorithm constructs CART models by minimizing an impurity function. On the other hand, the Bayesian approach uses many priors as the variable priors or the tree shape prior to construct trees with the maximum posterior probability. Once the trees are developed under these two approaches, the results are compared to determine which method gives the best trees.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

REMERCIEMENTS

Le tout premier remerciement revient à Jean-François Angers, mon directeur de maîtrise, sans qui ce projet n'aurait pu se réaliser. Sa patience et son dévouement m'ont grandement aidée alors que j'occupais un emploi tout en complétant ma maîtrise. Merci aussi à Sylvie Makhzoum pour tout le temps passé sur mon projet et pour ses encouragements. De plus, je voudrais remercier l'organisme MITACS qui m'a permis d'avoir un financement pour ce projet. Je tiens également à remercier Nadine Ouellette qui a beaucoup contribué à cette maîtrise en me supervisant tout au long de mon stage chez Meloche Monnex. Sa compétence et son professionnalisme ont été un atout très important à ma maîtrise. Merci à Éric Lacombe pour sa patience et pour la qualité du travail qu'il a accompli sur l'outil de data mining SAS Enterprise Miner. Aussi, je voudrais remercier Maureen Johnson pour sa contribution au projet.

Un remerciement tout spécial à Caroline qui m'a énormément donné de temps durant les derniers mois. En plus de son grand soutien, elle m'a pratiquement tout montré sur la programmation SAS.

Je tiens également à remercier mon copain Jean-François pour sa compréhension et sa patience légendaire. Je le remercie pour son soutien pendant toute cette période d'intense nervosité. J'aimerais aussi souligner l'appui de ma famille qui m'a toujours encouragée dans toutes mes entreprises. Mes parents et mon frère sont des modèles de persévérance et m'ont toujours poussée à aller plus loin. Merci aussi à ma belle-famille et à mes amis pour tout leur encouragement.

Enfin, je tiens à remercier tous les membres de l'équipe FBI de Meloche Monnex pour m'avoir épaulée quotidiennement. Je remercie tout particulièrement

Guillaume Gautier pour son soutien et sa compréhension. Merci aussi à Irshad Ruhomutually pour ses précieux conseils relativement à l'écriture en anglais.

INTRODUCTION

In the business and financial industries, data mining is becoming more and more important. It allows to perform analysis on large data sets and therefore identify key elements which describe the customers. For example, cluster analysis segments the database in order to have a better knowledge of the different groups that form the data set. Data mining is also used to forecast and predict specific customer characteristics with methods such as decision trees and neural networks. This Masters thesis presents two data mining techniques applied in the insurance business. These methods were developed during an internship at TD Meloche Monnex, a provider of group home and auto insurance for professionals and alumni. The project is done with the MITACS internship program which allows the collaboration between a partner organization and a university.

TD Meloche Monnex "Business Strategies" department requested this analysis since they wanted to gather further information relative to some aspects of the client population. Due to their demand, this Masters thesis is written in English. Their main objective is to understand a new category of the TD Meloche Monnex clients: the Direct Market individuals. These individuals do not belong to any TD Meloche Monnex group or association as for instance, professional or employer groups. This segment of clients is very different from the other groups of customers because it includes the general public. Hence, it is becoming increasingly more important to have a better understanding of these clients in order to identify which individuals are potential buyers.

To have a better knowledge of the population under consideration, an approach consists of dividing it in many homogeneous groups. Indeed the population itself is very heterogeneous so it is difficult to describe it entirely. Cluster analysis is thus used to create the groups by using different techniques. To see what is the best approach to answer this problematic, hierarchical and partitional clustering are studied. Furthermore, many distances and proximity measures are considered and the method giving the best result is therefore chosen. Once the groups (clusters) are created, they can be described and analyzed to see which one of them have desirable characteristics on a marketing and business point of view.

Once we have a better understanding of the population of Direct Market customers, the next step consists of predicting what kind of individual in this market is more likely to buy TD Meloche Monnex products or services. Hence, given that an individual asked for a quote, we wish to be able to predict if he will buy the product. The target variable used is the closing ratio which is the proportion of sales over the quotes. To develop a predictive model for this target variable, many statistical methods could have been used like logistic regression or neural networks. However, in the business context, decision trees allow to obtain a visual aspect of the model. Is is also easy to explain to business units.

In this Masters thesis, two approaches are used and compared to develop decision trees. First, the classical approach is considered to develop CART (classification and regression tree). To construct the model, it uses an algorithm presented by Breiman *et al.* (1984). This method defines a goodness of split function (the impurity) that must be minimized to find the best splitting rule.

The other method to construct decision trees is done under the Bayesian approach. This technique, called BART (Bayesian and regression tree) uses the *a priori* information to create a tree with the maximum posterior probability as possible. Indeed, the variable trends and frequencies are studied and incorporated in the *a priori* information. Furthermore, the prior distributions on variables depend on each variable importance from a business stand point. The desired tree characteristics like its shape or its number of terminal nodes are also included in the model prior. Therefore, the algorithm presented by Chipman and McCulloch (1998) and Denison and Mallick (2000) is developed to obtain BART models.

The Masters thesis starts with the explanation of the problematic and with definitions of insurance notions that are helpful in order to better understand the problematic. The database and the predictor variable are then presented followed by the descriptive analysis. This first chapter concludes with the explanation of a statistical notion applied in Chapter 3. The next chapter explains the cluster analysis and the decision tree model under the classical approach. To illustrate these concepts, a practical example is presented and used in Chapters 2 and 3. Finally, Chapter 3 exposes the BART approach by presenting all the elements used in the calculation of the tree posterior probability. It also explains the algorithm of the construction and the method to choose the best tree among many possibilities.

Chapter 1

INSURANCE AND STATISTICAL NOTIONS

This chapter begins with the presentation of the project objectives. To better understand these, some insurance notions are explained in Section 1.2. Then, the database and the variables used for the analysis are introduced in Section 1.3. A descriptive analysis is made in Section 1.4 and this chapter concludes with the explanation of a statistical method that is applied later in the project.

1.1. Objectives

This Masters thesis has two main objectives:

(1) to understand the Direct Market population,

(2) to model the closing ratio of this population.

1.1.1. Direct Market population

TD Meloche Monnex is a provider of group home and auto insurances for professionals and alumni. Hence, the TD Meloche Monnex clients are divided under three main segments due to some acquisitions. These are:

(1) employers and affiliated members,

(2) alumni and professional associations,

(3) direct market.

When Meloche Monnex was created, there was only the second segment consisting of student groups and also of university and professional associations. Later on, the TD Bank bought Meloche Monnex and the third segment that includes the Direct Market segment was created. Then, Meloche Monnex bought Canada Life and LICC, two insurance companies specialized in employer groups and affiliated members (the first segment).

Definition 1.1.1 (Direct Market). *The **Direct Market** population includes individuals that do not belong to any TD Meloche Monnex affinity group or association. The general public and TD clients are in this segment.*

Definition 1.1.2 (Affinity group). *A group is considered an **affinity group** if an agreement is or may be established with this group.*

Since Direct Market clients refer to the general public, their characteristics are not yet well known. The first objective of this project is thus to better understand this population by describing it and doing some segmentation.

1.1.2. Model the closing ratio

The second objective of this project is to know which individual in the Direct Market population is more likely to buy a TD Meloche Monnex insurance product. Based on some personal characteristics, we want to predict if an individual will buy a product given that we provide him with a quote. In this context, the variable to model is the closing ratio.

Definition 1.1.3 (Closing Ratio). *The **closing ratio** represents the proportion of sales among all the quotes for a given product. Indeed this ratio refers to the probability of buying an insurance product given a quote has been offered.*

Some models using the closing ratio as target variables must be constructed using many predictor variables. In this project, we use the total closing ratio that is for auto and residential products combined.

1.2. Definitions

Because the analysis is done in the insurance business, some insurance notions must be defined. Indeed the following definitions help to understand the database described later.

Definition 1.2.1 (Insurance). *Any individual is exposed to a significant amount of risk associated with perils like death, fire, disability, and so on. By purchasing an* ***insurance policy****, an individual transfers this risk to the insurance company (Brown and Robert, 2001).*

Definition 1.2.2 (Claim). *A* ***claim*** *is a demand for payment by an insured or by an injured third party under the terms and conditions of an insurance contract.*

An other important concept is the fiscal year. Indeed most variables are calculated at the end of the fiscal year instead of the calendar year.

Definition 1.2.3 (Fiscal year). *The* ***fiscal year*** *begins at the November month of preceding year. For example, year 2007 is from November 2006 to October 2007.*

Definition 1.2.4 (Fiscal month). *The* ***fiscal month*** *differs from the calendar month in that it generally ends on the last Friday of the month.*

As we will see in Section 1.3, the data set includes two categories of individuals: the actual clients and the prospects.

Definition 1.2.5 (Prospects). *The* ***prospects*** *are individuals that the company can potentially have as clients.*

Definition 1.2.6 (Expiration Date). *The* ***expiration date*** *is the date after which the insurance policy is no longer valid.*

Every individual in the data set has an entry for the expiration date. Furthermore, this date either refers to the TD Meloche Monnex policy (for the actual clients) or for another insurer policy (for the prospects). In order to convince the prospects to buy TD Meloche Monnex insurance products, those potential clients are called forty five days before their expiration date. The list of prospects is obtained by three ways:

(1) individuals who used to be TD Meloche Monnex clients,
(2) individual who called in on TD Meloche Monnex for a quote,
(3) individuals who have been targeted and called by the telemarketing department in order to obtain their expiry dates. For instance, a marketing campaign may target students from a specific university by doing some promotions. The targeted clients will be called and considered as prospects.

Furthermore, the following definitions explain some concepts in the insurance policy coverage.

Definition 1.2.7 (Collision coverage). *When a policy has the **collision protection**, if the vehicle is damaged in an accident, the insurer will pay the cost of its repair or replacement as defined in the policy.*

Definition 1.2.8 (Comprehensive coverage). *The **comprehensive protection** covers repairs on a damaged vehicle due to a peril other than collision such as fire, vandalism, stone chips and so on.*

Definition 1.2.9 (Deductible). *The amount of **deductible**, let say d, means that the policyholder is responsible for the first \$d of the repair or replacement cost. This tends to eliminate the filing of small claims for which the cost of administration and settlement would likely exceed the benefits (Brown and Robert,2001).*

1.3. Variables and database

The data set consists of all the Direct Market clients and prospects in 2005 for the four regions of Canada (Québec, Ontario, Western Provinces, Atlantic Provinces). Each entry in the data set refers to a client or a prospect. One client or prospect may have several automobile or residential policies. Furthermore, each automobile policy can include more than one vehicle and each residential policy can cover many homes.

Because TD Meloche Monnex is an insurer for automobile and residential products, the studied variables are divided in the following categories:

(1) the demographic variables,
(2) the auto variables,
(3) the residential variables.

1.3.1. Demographic variables

The demographic variables are variables that describe the individuals with characteristics other than their home and residential policy characteristics. The demographic variables present in the database are:

- **Account since**: number of months since the first quote was made on the account,
- **Gender**: gender of the account's principal owner,
- **Average income**: household average income viewed at the end of fiscal year.

1.3.2. Automobile variables

The individuals in the data set that have available information on automobile variables are those who own an auto policy or prospects who had a quote made for this kind of policy. The prospects that only made a residential quote do not have the auto characteristic and therefore have missing values. The automobile variables are:

- **Driving record**: the number of years since the last accident on a given account,
- **Creditor**: variable that indicates if there is a creditor on at least one of the vehicles on the account at the end of the fiscal year,

- **Renting**: indicates if there is a renting agreement on at least one of the vehicles on the account at the moment of the renewal,
- **High performance vehicle**: indicates if there is at least one high performance vehicles on the account at the moment of the renewal,
- **Vehicle deductible**: refers to the sum of each vehicle deductible amount on the account,
- **Motorcycle**: the number of motorcycles in the account in the last year viewed at the end of the fiscal year,
- **Private passenger vehicle (PPA)**: the number of private passenger vehicles in the account in the last year viewed at the end of the fiscal year,
- **All-Terrain Vehicle (ATV)**: the number of all-terrain vehicles in the account in the last year viewed at the end of the fiscal year,
- **Snowmobile vehicle**: the number of snowmobiles in the account in the last year viewed at the end of the fiscal year,
- **Other vehicle**: the number of other vehicles such as trailers, vintage or motorhomes in the account in the last year viewed at the end of the fiscal year,
- **Sales**: the number of sales by client or prospect in the last year viewed at the end of the fiscal year,
- **Quotes**: the number of quotes by client or prospect in the last year viewed at the end of the fiscal year,
- **Responsible claim**: the number of responsible active collision claim files in the last 3 years, viewed at the end of the fiscal year,
- **Non responsible claim**: the number of non responsible active collision claims filled in the last three years, viewed at the end of the fiscal year,
- **Comprehensive claims**: the number of comprehensive active claim files in the last three years, viewed at the end of the fiscal year,
- **License since**: the number of months since the youngest client on the account has his driver license,

- **Collision coverage**: indicates if there is a collision coverage on the account,
- **Vehicle age**: the age of the oldest active vehicle of the policy viewed at the end of the fiscal year.

1.3.3. Residential variables

The residential variables are:

- **Homeowner package**: number of homeowner packages in the account in the last year viewed at the end of the fiscal year,
- **Condo package**: number of condo packages in the account in the last year viewed at the end of the fiscal year,
- **Tenant package**: number of tenant packages in the account in the last year viewed at the end of the fiscal year.

1.4. Descriptive analysis

As we explained above, the analysis is done on the Direct Market clients and prospects in the four regions of Canada using 2005 data. It is interesting to see how this population is distributed among the four regions of Canada. Table 1.1 shows that the majority of the Direct Market portfolio is in the province of Ontario. There is a similar number of observations in the regions of Québec and Western and a small number in the region of Atlantic.

Tab. 1.1. Direct Market population distributed by region in 2005.

Region	Percentage of observations (%)
Ontario	61
Québec	19
Western	17
Atlantic	2

1.4.1. Descriptive analysis for Ontario

Since the majority of the Direct Market clients and prospects are in the province of Ontario, the descriptive analysis is presented for this province only.

The data set includes 153,028 clients and prospects in Ontario and 34 variables. However, only the most significant variables are described.

1.4.1.1. *Demographic variables*

The demographic variables are variables that describe the individuals without regard to their automobile or residential information. Consequently, these variables are available for all the observations in the data set. Table 1.2 shows that there is an important majority of males in the population and that clients and prospects are almost equally represented.

TAB. 1.2. Percentages of observations for the variables "gender" and "clients" for the province of Ontario in 2005.

Characteristic	Percentage of observations (%)
Male	72
Client	46

For the continuous variables, some descriptive statistics are presented in Table 1.3. The average age is around 40 and 28% of the population buys an insurance product. Furthermore, the distribution of the variable "age" is represented in Figure 1.1.

1.4.1.2. *Automobile variables*

This subsection describes auto characteristics for individuals with at least one auto policy (clients) or who asked for an auto quote (prospect). In Ontario, 89% of the Direct Market population in 2005 had this characteristic. The 11% left are individuals that have only residential products. Table 1.4 shows the automobile

TAB. 1.3. Descriptive statistics for continuous demographic variables for the province of Ontario in 2005.

Variable	Mean	Standard deviation	Minimum	Maximum
Age	40	12	16	90
Household average income	68,495	29,356	9,091	928,844
Closing ratio (%)	28	43	0	100

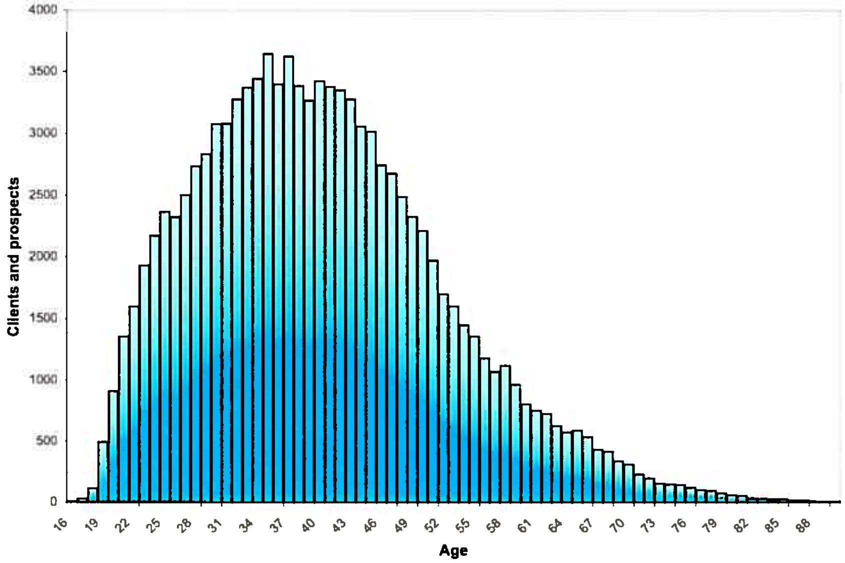


FIG. 1.1. Histogram of the variable "age".

characteristics for this subpopulation. For example, among the 89%, 65% of the individuals have or asks for collision protection. Note that the sum of all the percentages is not equal to 100% because a client or a prospect can have more than one of these characteristics. For instance, a client may have a moto vehicle and a collision protection on his vehicle.

TAB. 1.4. Percentages of observations with each auto characteristic for the province of Ontario in 2005.

Characteristic	Percentage of observations (%)
Collision protection	65
Private passenger vehicle	88
Snowmobile vehicle	9
Moto vehicle	3

Table 1.5 presents other auto characteristics. These continuous variables are based on time notion and they all have a minimum value of 0. For example, the minimum of 0 for the vehicle age means that the vehicle was new at the moment of the creation of the data set.

TAB. 1.5. Descriptive statistics for continuous variables related to auto policy for Ontario.

Variable	Mean	Standard deviation	Minimum	Maximum
Vehicle age	7	6	0	51
Licence since (in months)	165	150	0	895
Account since (in months)	22	22	0	140

1.4.1.3. *Residential variables*

The characteristics described in this subsection are for individuals with at least one residential policy (clients) or who asked for a residential policy (prospect). In Ontario, 25% of the Direct Market population in 2005 had this characteristic. The 75% left are individuals that have only automobile products. In Table 1.6, we see that among the 25%, 69% of the observations have or asked for a homeowner package.

TAB. 1.6. Percentages of observations with each characteristic in Ontario.

Characteristic	Percentage of observations (%)
Homeowner package	69
Condo package	12
Tenant package	19

1.4.2. Average closing ratio

To model the closing ratio, only individuals who made a quote are included in the database. Indeed when the number of quotes is null, the closing ratio is missing. Consequently, the number of observations in the database is reduced so that each person has a value for the closing ratio. Table 1.7 shows the number of observations in each region with the corresponding closing ratio.

TAB. 1.7. Number of observations and closing ratio for each region.

Region	Number of observations	Closing ratio (%)
Ontario	105,635	28
Québec	43,791	15
Western	31,666	25
Atlantic	4,074	27

In Table 1.7, we see that the regions of Ontario, Western and Atlantic have a similar closing ratio compared to Québec which has the smallest among the different regions. This is because the Québec market is more competitive and it is not as much regulated as the other provinces.

1.5. EM ALGORITHM FOR A MIXTURE MODEL

This section explains the EM algorithm that is used later in this project. In our context, this algorithm is used to estimate the parameters of a mixture model. Furthermore, as we will see in Chapter 3, the mixtures used here are mixtures of only two distributions.

Suppose that the graphic representation of a variable is given by the Figure 1.2. This bimodal graphic indicates that the variable has a mixture distribution.

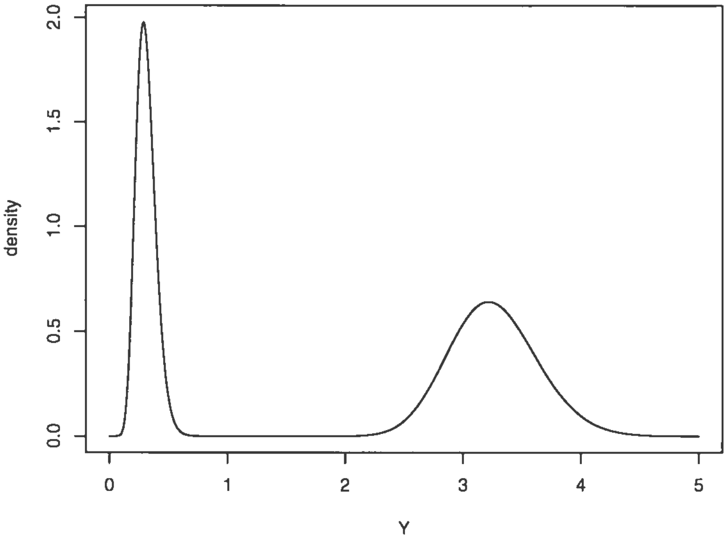


FIG. 1.2. Density of a mixture of two normal variables.

We can define the random variable Y having a mixture of two distributions as:

$$Y_1 \sim \phi_{\theta_1}(y),$$

$$Y_2 \sim \phi_{\theta_2}(y),$$

$$Y = (1-\Delta)Y_1 + \Delta Y_2,$$

where $\Delta \in \{0,1\}$ with $\mathbb{P}(\Delta = 1) = p$. The density of y is therefore

$$f_Y(y) = (1-p)\phi_{\theta_1}(y) + p\phi_{\theta_2}(y),$$

where p is the probability that an observation follows the distribution ϕ_{θ_2}.

To estimate the parameters p, θ_1 and θ_2, the log-likelihood is calculated as:

$$l(\theta; Z) = \sum_{i=1}^{N} \log[(1-p)\phi_{\theta_1}(y_i) + p\phi_{\theta_2}(y_i)].$$

To maximize $l(\theta; Z)$, an iterative method is used and many iterations are required before convergence. In this Masters thesis, R project was used to execute these iterations. The algorithm is:

(1) take initial values for the parameters p, θ_1 and θ_2,

(2) compute the responsibilities:

$$\zeta_i = \frac{p\phi_{\theta_2}(y_i)}{(1-p)\phi_{\theta_1}(y_i) + p\phi_{\theta_2}(y_i)},$$

(3) generate $u \sim$ Bernouilli(ζ_i),

(4) minimize the log-likelihood versus θ_1 and θ_2:

$$l_0(\theta; Z, u) = \sum_{i=1}^{N}[(1-u_i)\log\phi_{\theta_1}(y_i) + u_i \log\phi_{\theta_2}(y_i)] + \sum_{i=1}^{N}[(1-u_i)\log\hat{p} + u_i \log\hat{p}],$$

where $\hat{p} = \frac{1}{N}\sum_{i=1}^{N}\zeta_i$ is the weighted parameters.

(5) Iterate steps 2 to 4 by replacing θ_1 by $\hat{\theta_1}$, θ_2 by $\hat{\theta_2}$ and p by $\hat{p}$ until convergence.

A way to choose the initial values of θ_1 and θ_2 is to take two y_i at random. For the parameter p, the initial value can be any value between 0 and 1. In this Masters thesis, a value of 0.5 was chosen for this initial parameter (see Hastie *et al.*, 2001).

This chapter began with the presentation of the two aims of the project. Also, it presented the variables included in the database. Then, the variables explained in Section 1.3 were used to describe the Direct Market population and to construct a model that predict the closing ratio. Chapter 2 will explain two classic statistical methods to resolve the two objectives.

Chapter 2

CLASSICAL APPROACH

In Chapter 1, the data set used for this project has been presented. This chapter describes two statistical methods applied to analyze these data. The main objective of this project is to predict who in the Direct Market segment is more likely to buy insurance products. However, in order to create statistical models, the understanding of the population under consideration is essential. Therefore, this chapter begins with the description of an exploratory data analysis method. Indeed, Section 2.1 discusses about clustering, a statistical method to explore and to classify the data. In order to describe Direct Market clients and prospects, this technique tries to form homogeneous sub-groups which are very different from each other. In this first section, the closing ratio is not yet modeled because the clustering is a descriptive method, *i.e.* a method that does not need a target variable.

Once the population of interest has been studied, the objective is to find which individuals are more likely to buy insurance products. Hence, a statistical model must be developed to find these individuals. Therefore, some decision trees are constructed using the closing ratio as target variable. This is described in Section 2.2.

2.1. Clustering

An objective of this project is getting to know a particular segment of Meloche Monnex prospects and clients that is, the direct market. However, because the population in this segment is composed of the general public, many different individuals are in it. The population is thus an heterogeneous population. Therefore, describing it globally could be misleading. The cluster analysis answers this problem. This method has the purpose of grouping clients and prospects into groups or clusters based on similarity in their characteristics.

This section begins with the description and the definition of the cluster analysis. Then, we discuss the data preparation in Section 2.1.2. Afterward, some similarity and distance measures are presented in Sections 2.1.3 and 2.1.4. We follow in Section 2.1.5 with the explanation of two clustering strategies: the Hierarchical clustering and the Partitional clustering. To conclude this section, we elaborate about the method used to find the optimal number of clusters.

2.1.1. Description

Cluster analysis is a common technique in exploratory data analysis. It is the classification of objects into different groups. More precisely, the data set is separated into subsets or clusters, so that all the objects in each cluster tend to be similar to each other. There is many way to cluster a data set. Therefore, the underlying mathematics of most of these methods are relatively simple but large numbers of calculations are needed.

Definition 2.1.1 (Cluster)**.** *A **cluster** is a group of contiguous elements of a statistical population; for example, a group of people living in a single house, a consecutive run of observations in an ordered series, or a set of adjacent plots in one part of a field (*cf. *Everitt, 1993).*

Definition 2.1.2 (Good clusters)**.** ***Good clusters** are clusters that present little variation into the groups and large variation between the groups. They also need to be large enough to be significant.*

We begin with an example of a data set that could be divided into clusters.

Example 2.1.1. *This simulated data set is composed of 10 individuals with the following characteristics: their age, their number of auto claims and their closing ratio. It is represented in Table 2.1. The cluster analysis is done using "age" and "number of auto claims". If an observation does not have any auto policy, its number of auto claim is missing ("."). With this dataset, it is possible to group some similar individuals and produce 2 differents clusters. One group could include young persons with many auto claims while the other could be formed of older individuals with a smaller number of claims.*

TAB. 2.1. Data of Example 2.1.1.

Observations	*Age*	*Number of auto claims*	*Closing ratio*
1	*30*	*0*	*1.00*
2	*42*	*1*	*0.30*
3	*20*	*2*	*0.75*
4	*28*	*3*	*0.00*
5	*55*	*1*	*0.30*
6	*33*	*4*	*0.40*
7	*35*	*3*	*0.40*
8	*30*	*5*	*0.40*
9	*44*	*.*	*0.00*
10	*51*	*.*	*0.25*

2.1.2. Preparing the data

In a large database with many variables, the data must be preprocessed before they are analyzed. First of all, the missing values must be examined carefully. Indeed, missing values can have different meanings depending on the variable. As we explained in Chapter 1, the variables can be demographic, automobile or residential. For the demographic variables, the data set is cleaned in order to obtain no missing value. For the auto variables, we must only have missing values for individuals without an auto policy. The same principle is applied to residential variables while missing values must be for individuals without residential policy.

Another important consideration is the variable variances. For example, in the data set, some variables are expressed in thousands while others are in hundreds. Therefore, variables with large variances tend to have more effect on the resulting clusters than variables with small variances. It is thus recommended to standardize these variables. However, if all variables are measured in the same units, there is no need for standardization.

Definition 2.1.3 (standardization). *A variable x_i is **standardized** when its values are transformed and given by:*

$$\frac{x_{ij} - \mu_i}{\sigma_i},$$

where μ_i is the mean of x_i and σ_i, his standard deviation.

2.1.3. Similarity measure between individuals

A clustering method attempts to group the objects based on some measures of similarity. Similarities are a set of rules that serve as criteria for grouping or separating items. It is possible to measure similarity and dissimilarity in a number of ways. Consequently there is not a single correct classification. In order to measure the similarity, an important concept is the similarity matrix. This matrix represents the similarities or the dissimilarities between the individuals present in the data set. It is used for the clustering algorithms. Therefore, we note D, the similarity matrix that is a $n \times n$ matrix where n is the number of observations. Each element of this matrix, noted $d_{jj'}$ is the similarity between the j^{th} and the j'^{th} observation. This matrix is also symmetric and the diagonal elements are null.

To compute the matrix D, a similarity measure must be specified. Thus, it is more common to measure the similarity as the dissimilarity between objects. Therefore, we define $x_1, ...x_p$ the predictor variables and $d(x_{ij}, x_{ij'})$, the dissimilarity measure between x_{ij} and $x_{ij'}$, the values of the predictor i for the observations j and j'. The dissimilarity between individuals j and j' is therefore function of $d(x_{ij}, x_{ij'})$, where $i = 1, \ldots, p$, $j = 1, \ldots, n$, $j' = 1, \ldots, n$ and p is the number of predictor variables. The value of $d(x_{ij}, x_{ij'})$ can be determined by many different functions. These functions depends on the variable types that can be quantitative, ordinal or binary.

2.1.3.1. *Similarity measure for quantitative variables*

For the quantitative variables, we present the two most important measures:

(1) **Euclidean distances**: this is the most commonly chosen type of distance. It is the geometric distance in the multidimensional space. The distance between x_{ij} and $x_{ij'}$ is computed as:

$$d(x_{ij}, x_{ij'}) = (x_{ij} - x_{ij'})^2 \text{ and}$$
$$D(x_j, x_{j'}) = \sqrt{\textstyle\sum_{i=1}^{p} d(x_{ij}, x_{ij'})}.$$

(2) **Manhattan distance**: this distance is the average absolute value difference across dimensions.

$$d(x_{ij}, x_{ij'}) = |x_{ij} - x_{ij'}| \text{ and}$$
$$D(x_j, x_{j'}) = \textstyle\sum_{i=1}^{p} d(x_{ij}, x_{ij'}).$$

In Hastie *et al.* (2001), a similarity measure based on the correlation between variables is described. In this case, the similarity measure is a similarity measure and is :

$$\rho(x_j, x_{j'}) = \frac{\sum_i (x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'})}{\sqrt{\sum_i (x_{ij} - \overline{x}_j)^2 \sum_i (x_{ij'} - \overline{x}_{j'})^2}} \quad ,$$

where $\overline{x}_{j'} = \frac{1}{p}\sum_{i=1}^{p} x_{ij}$ is the average for the observation j over the p variables.

2.1.3.2. *Similarity measure for ordinal variables*

Another consideration is the distances measures for the ordinal variables. These variables are those where all possible values are ranked depending on their importance. In that situation, the variables are transformed before the computation of the dissimilarity matrix. The measure is given by (*cf.* Hastie *et al.*, 2001):

$$x'_{ij} = \frac{x_{ij} - 1/2}{M},$$

where x_{ij} is the value of the observation j for the variable i such as $x_{ij} = 1, ..., M$ and M is the number of categories for this variable. For each ordinal variable, this measure replaces the original one and the similarity matrix could be calculated using this transformed value. They are then treated as quantitative variables.

2.1.3.3. *Similarity measure for binary variables*

Many binary variables, also called dichotomous variables, are included in our data set. They categorize data in two groups with value 0 for one group and 1 for the other group. Suppose we observe the contingency table given in Table 2.2. where n is the number of observations.

TAB. 2.2. Values to calculate similarity measure for dichotomous variables.

$x_{ij'} \backslash x_{ij}$	0	1	Total
0	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+d$	n

Therefore, we can now define different dissimilarities measures (*cf.* Lorr, 1983). The usual dissimilarity functions are:

(1) $d(x_{ij}, x_{ij'}) = \frac{a+d}{n}$ (Coefficient of concordance),

(2) $d(x_{ij}, x_{ij'}) = \frac{d}{b+c+d}$ (Jacquard coefficient),

(3) $d(x_{ij}, x_{ij'}) = \frac{2d}{2d+b+c}$,

(4) $d(x_{ij}, x_{ij'}) = \frac{2(a+d)}{2(a+d)+b+c}$,

(5) $d(x_{ij}, x_{ij'}) = \frac{d}{d+2(b+c)}$.

Now that some distances have been defined, the dissimilarities between x_j and $x_{j'}$ is given by:

$$D(x_j, x_{j'}) = \sum_{i=1}^{p} d(x_{ij}, x_{ij'}),$$

where p is the number of predictor variables. Therefore, since the similarity matrix can be determined, we have all pairwise distances for the individuals.

2.1.4. Distance measure between clusters

Now that the similarity between all individuals is determined, the distance between clusters can be computed. However, because the clusters include many individuals, the distance between clusters is not easily calculated. Therefore, many distance functions exist. The usual measures to calculate the distance between these two clusters are:

Centroid distance: the distance between groups is the distance between the cluster centers called the group centroids. Therefore, in order to find the distance between two groups, let say H and L, the cluster centers $\overline{m}_H$ and $\overline{m}_L$ must be determined. The centroid distance between these two clusters is:

$$d^{\text{centroid}}(H, L) = |\overline{m}_H - \overline{m}_L|,$$

where $\overline{m}_H = \frac{1}{p}\frac{1}{n_H}\sum_{j=1}^{p}\sum_{h=1}^{n_H} x_{jh}$, $\overline{m}_L = \frac{1}{p}\frac{1}{n_L}\sum_{j=1}^{p}\sum_{h=1}^{n_L} x_{jh}$, n_H and n_L are the numbers of observations in groups H and L respectively.

This measure is not appropriate when the sizes of the two clusters to be grouped are very different. In this case, the centroid of the new group will be very close to the centroid of the larger group. Thus, the properties of the smaller group are then virtually lost (*cf.* Everitt, 1993). However, this measure has the advantage of only having to calculate the difference between each cluster centroid. In opposition, the three other distances described below need the calculation of the differences between every pairs of individuals in the two groups.

Single linkage clustering: this measure is also called "Minimum or Nearest-Neighbour Method". The dissimilarity between 2 clusters is the minimum dissimilarity between members of the two clusters, that is

$$d^{\text{single}}(H, L) = \min_{x_h \in H, x_l \in L} D(x_h, x_l).$$

This measure has the advantage of being the simplest but has the disadvantage that an outlier can cause two groups of individuals to be clustered when most of the individuals are really distant.

Complete linkage clustering: this measure is also called "Maximum or Furthest-Neighbour Method". The dissimilarity between 2 groups is equal to the greatest dissimilarity between a member of a given cluster and a member of the other one. This method tends to produce very tight clusters of similar cases. The distance between clusters H and L is :

$$d^{\text{complete}}(H, L) = \max_{x_h \in H, x_l \in L} D(x_h, x_l).$$

Complete linkage has the advantage over single linkage in that within a cluster, all pairs of individuals will be within the distance at which the cluster was formed.

Group Average Method: the distance between groups is the average of the distances between pairs of individuals in the two groups, that is, the distance between clusters H and L is:

$$d^{\text{average}}(H, L) = \frac{1}{n_H n_L} \sum_{h=1}^{n_H} \sum_{l=1}^{n_L} D(x_h, x_l).$$

This measure is a good compromise between the extremes of single and complete linkage, but the distances at which clusters are formed are averages, not real distances. Therefore, the clusters can be more difficult to interpret. However, it takes longer to evaluate.

2.1.5. Clustering strategies

Two main clustering strategies are discussed in this chapter:

(1) hierarchical clustering,

(2) partitional clustering.

2.1.5.1. *Hierarchical clustering*

In hierarchical clustering, the data set is not partitioned into a particular cluster. Instead, a series of partitions or merges take place, which may run from a single cluster containing all objects to n clusters each containing a single object or the other way around, that is from n clusters to 1. Hierarchical clustering techniques are subdivided into top-down and bottom-up methods. A top-down method begins with all observations in the same cluster. This cluster is gradually broken down into smaller and smaller clusters. Bottom-up techniques are more commonly used. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively. At each stage, the two most similar clusters are joined and it continues until there is just a single cluster. Therefore, the single observations are the smaller clusters possible. Hierarchical clustering may be represented by a two dimensional diagram known as a dendrogram (see Figure 2.2) which illustrates the fusions or divisions made at each successive stage of the analysis.

Definition 2.1.4 (Dendrogram)**.** *A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm. The vertical axis represents the distances between clusters and the horizontal axis is the observation sequence numbers. Each vertical line represents a cluster.*

Hierarchical clustering (see section 2.1.5.1) was applied to Example 2.1.1. Figure 2.1 shows three dendrograms using Euclidean distance and three distance measures between clusters. Although the distances between clusters are different, it is possible to see that the merges are similar for all measures.

The process of bottom-up hierarchical clustering can be summarized as follows:

(1) calculate the distance between all initial clusters. In most analysis, initial clusters will be made up of individual cases,

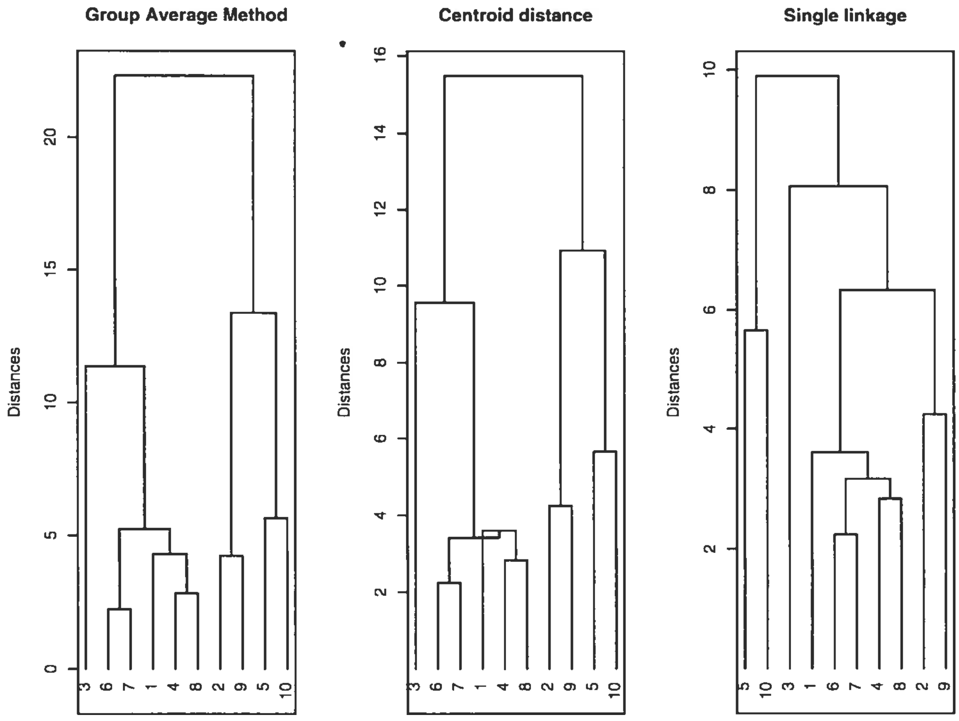


FIG. 2.1. Dendrogram produced by hierarchical clustering on Example 2.1.1 using Euclidean distance and three distance measures. The final clusters are similar for the three distance methods.

(2) merge the two most similar clusters and recalculate the distances,

(3) repeat step 2 until all cases are in the same cluster.

The data set presented in Example 2.1.1 can be classified using bottom-up clustering (see Figure 2.2). It is possible to see that the clustering algorithm begins with each observation being a single cluster. Then, we see that individuals 6 and 7 are similar so they are grouped together. Table 2.3 shows the nine merging steps. Thus, at step 7, there is three clusters formed with (1, 3, 4, 6, 7, 8),(2, 9) and (5, 10).

In order to choose the final clusters, the dendrogram is used. Therefore, this representation shows the distance between the merged clusters. The more this

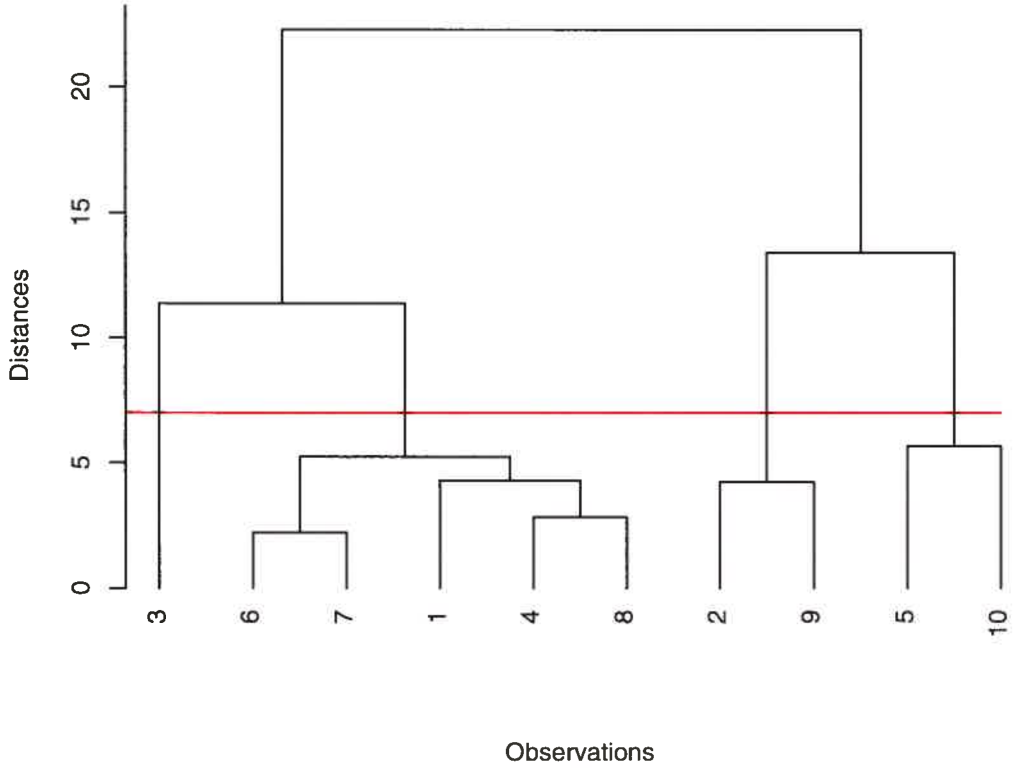


FIG. 2.2. Dendrogram produced by hierarchical clustering using Euclidean distance and group average method. It shows the different cluster merges. At the beginning, each observation forms a single cluster. Then, the observations are grouped until there is just one cluster. A visual inspection is used to choose the number of clusters by cutting the dendrogram at the desired level. The red line represents the cut off distance where the merges are stopped.

distance is important, the more clusters are different from another. As the clusters are merged, the distances between them increase until the clusters are too dissimilar. At this moment, the clusters must stay distinct and the merging process stops. Thus, an horizontal line is drawn in the dendrogram at this distance. The number of vertical lines that cross the horizontal line corresponds to the correct number of clusters. In Figure 2.2, the red line shows that there are four

TAB. 2.3. Merging process of the hierarchical clustering in Example 2.1.1.

Steps	Clusters
1	(1),(2),(3),(4),(5),(6,7),(8),(9),(10)
2	(1),(2),(3),(4,8),(5).(6.7),(9),(10)
3	(1,4,8),(2),(3),(5),(6,7),(9),(10)
4	(1,4,8),(2,9).(3),(5),(6,7).(10)
5	(1,4,6,7,8),(2,9),(3),(5),(10)
6	(1,4,6,7,8),(2,9).(3),(5,10)
7	(1,3,4,6,7,8),(2,9),(5,10)
8	(1,3,4,6,7,8),(2,9.5,10)
9	(1,3,4,6,7,8,2,9.5,10)

different clusters. The observation 3 forms a single cluster, the second includes (1, 4. 6, 7. 8). and the clusters 3 and 4 are respectively formed with (2, 9) and (5. 10).

In hierarchical clustering, there is a particular merging method called Ward's classification. According to Ward (1963), the loss of information which results from grouping two clusters can be measured by the total sum of squared deviations. At each step, the union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the error of squares are combined (Everitt. 1993).

Hierarchical clustering is easily calculated but it is not adapted for large data sets. Furthermore, it does not allow provision for reallocation of entities who may have been poorly classified at an early stage in the analysis.

2.1.5.2. *Partitional clustering with K-means clustering*

Partitional clustering is a clustering method that directly divides the data set into clusters. The clustering algorithm optimizes a criterion function based on two restrictions. It must minimize some measure of dissimilarity within the clusters and must maximize the dissimilarity between the different clusters.

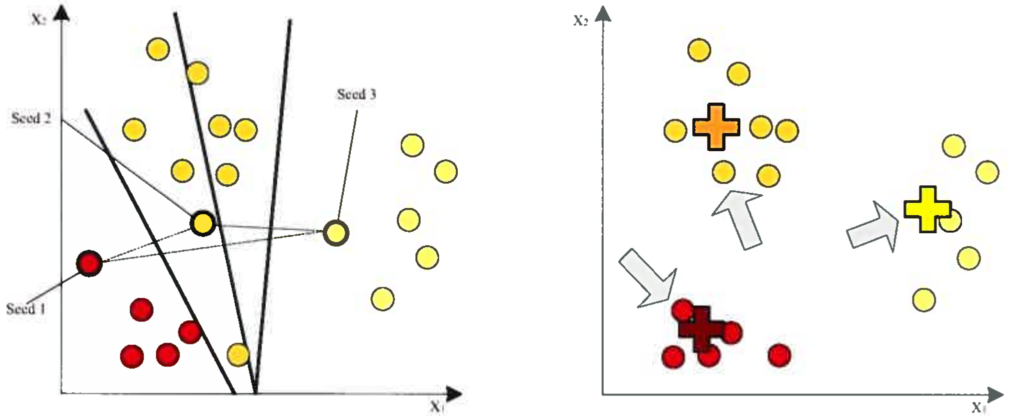


FIG. 2.3. Illustration of the K-means algorithm.

Partitional clustering differs from hierarchical clustering in that they admit relocation of the observations. Therefore, a poor classification might be corrected at a later stage. Hence, this is the chosen method for this project because it is appropriate for the efficient representation and compression of large databases. The partitional technique presented here is called the K-means method because it forms K different clusters. This method assumes that the number of groups has been decided *a priori*. The algorithm works as follow:

(1) randomly selects K seeds used as initial estimates of cluster centers. Many initialization methods can be used. For example, MacQueen (1967): chooses the first K points in the sample as the initial cluster mean vectors.

(2) Assign each record to its closest cluster center.

(3) Compute new cluster centers as the centroids of the clusters.

(4) For each observation, calculate its distance from each centroid.

(5) Repeat step 2 to 4 until convergence.

This algorithm is guaranteed to converge (see Andersberg, 1973).

2.1.6. Number of clusters in K-means

One of the most important problem of cluster analysis is to identify the optimum number of clusters. Consequently, many methods have been developed to determine this number.
Caliński and Harabasz (1974) developed a ratio given by:

$$\text{ratio} = \frac{(n-K)\text{trace}(BSS)}{(K-1)\text{trace}(WSS)},$$

where n is the total number of observations, K is the number of clusters, WSS is the sum of squares within cluster, and BSS is the sum of squares between clusters.

Also, Duda and Hart(1973) proposed a criterion function that expresses how well a given K-cluster description matches the data. We expect a description in terms of $K+1$ clusters to give a better fit than a description in term of K clusters. Therefore, to see if there is a statistically significant improvement in having $K+1$ clusters instead of K clusters, the following ratio is computed:

$$\text{ratio} = \frac{WSS(K+1)}{WSS(K)},$$

where $WSS(\mathrm{K}+1)$ is the sum of squared errors within cluster when there is $K+1$ clusters and $WSS(K)$ is the sum of squared errors within cluster when there is a K cluster. The null hypothesis that there are exactly K clusters is rejected at the d-percent significance level if:

$$\frac{WSS(K+1)}{WSS(K)} < 1 - \frac{2}{\pi p} - \alpha\sqrt{\frac{2(1-8\pi^2 p)}{np}},$$

where p is the number of variables and α is such as $d = 1 - \Phi(\alpha)$.

Another criterion is proposed by Edwards and Cavalli-Sforza(1965). This approach minimizes the variability within groups as measured by the sum of the variation on each variable.

To estimate the number of clusters, the chosen criterion for this project is the Cubic Clustering Criterion(CCC). This criterion is provided by the SAS programming package (Sarle, 1983). It also minimizes the variability within groups. We begin with some notations:

WSS:= within-cluster sum of squares,
ESS:= error sum of squares,
n:= number of observations in the data set,
n_k:= number of observations in the k^{th} cluster,
p:= number of variables,
K:= number of clusters,
$X : n \times K$ matrix of variable observations,
$\overline{X} : K \times p$ matrix of cluster means,
$Z : n \times p$ matrix of cluster indicator with elements z_{ik} for which:

$$z_{ik} = \begin{cases} 1 & \text{if} \quad \text{the } i^{\text{th}} \text{ observation belongs to the } k^{\text{th}} \text{ cluster,} \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

Let, ZZ^t, a $K \times K$ diagonal matrix with the n_k on the diagonal and $k = 1, ..., K$, such that

$$\overline{X} = (\mathrm{ZZ^t})^{-1}\mathrm{Z^tX}.$$

The total-sample sum of squares and cross products (SSCP) matrix, denoted T is given by:

$$T = X^t X.$$

The between-cluster SSCP matrix BSS is:

$$BSS = \overline{\mathrm{X}}\mathrm{Z^tZ}\overline{\mathrm{X}}.$$

The within-cluster SSCP matrix is

$$\begin{aligned} WSS &= (\mathrm{X} - \mathrm{Z}\overline{\mathrm{X}})'(\mathrm{X} - \mathrm{Z}\overline{\mathrm{X}}) \\ &= \mathrm{X^tX} - \overline{\mathrm{X}}'\mathrm{Z'Z}\overline{\mathrm{X}} \\ &= \mathrm{T} - \mathrm{B}. \end{aligned}$$

The within-cluster sum of squares pooled over variables corresponds to the trace of WSS. Since T is constant for a given sample, minimizing trace(WSS) is equivalent to maximizing:

$$R^2 = 1 - \frac{\text{trace}(WSS)}{\text{trace}(T)}, \tag{2.1.1}$$

where R^2 is the proportion of variance accounted for by the clusters. The expected value of R^2, $E(R^2)$ is determined by the assumption that the data have been sampled from a uniform distribution based on a hyperbox. Therefore, in order to obtain an approximation of $E(R^2)$, we have to find an approximation for R^2 (*cf.* Sarle, 1983). The volume of the hyperbox, noted v is given by:

$$v = \prod_{j=1}^{p} s_j,$$

where s_j is the edge length of the hyperbox. If the hyperbox is divided into q hypercubes with edge length c, this length is given by:

$$c = \left(\frac{v}{q}\right)^{\frac{1}{p}}.$$

The number of hypercubes along the j^{th} dimension of the hyperbox is:

$$u_j = \frac{s_j}{c}.$$

Furthermore, we have that the total variance along the j^{th} dimension is proportional to ${s_j}^2$ and the within-cluster variance is proportional to c^2. Therefore, R^2 can be expressed by:

$$R^2 \;=\; 1 - \frac{\sum_{j=1}^{p} c^2}{\sum_{j=1}^{p} s_j^2}.$$

In Sarle (1983), they explain that the expected value $E(R^2)$, found with simulations is approximated by:

$$E(R^2) = 1 - \left[\frac{\sum_{j=1}^{p^*} \frac{1}{n+u_j} + \sum_{j=p^*+1}^{p} \frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p} u_j^2}\right] \frac{(n-q)^2}{n}(1 + \frac{4}{n}), \tag{2.1.2}$$

where p^* is an estimate of the dimensionality of the between cluster variation.

The CCC is estimated from the observed R^2 as:

$$CCC = \log\left[\frac{1 - E(R^2)}{1 - R^2}\right] \frac{\sqrt{\frac{np^*}{2}}}{(0.001 + E(R^2))^{1.2}}. \tag{2.1.3}$$

To estimate the number of clusters, the CCC is plotted against the number of clusters and the following conclusions can be made:

- the number of clusters that corresponds to maximums on the plot with the $CCC > 2$ or 3 is chosen and indicates good clustering.
- If there is a maximum with the CCC between 0 and 2, there is possible clusters but they should be interpreted cautiously.
- CCC is not appropriate for clusters that are highly elongated or irregularly shaped.

2.2. Decision trees

In this section, a statistical model is developed to find the individuals who are insurance buyers. Therefore, some statistical models called CART are produced. CART stands for Classification and Regression Trees. As the name implies, the CART methodology involves using trees to resolve classification and regression problems.

This section starts with the description of the CART structure. Then, the method to construct the trees is explained in Section 2.2.2. To better understand this aspect, Section 2.2.3 explains the set of questions used to split the tree. Section 2.2.4 discusses about some particularity of the data that affects the method used for this analysis. In Section 2.2.5, the measures of goodness of split are presented and illustrated by means of an example. Furthermore, to determine how large to grow the tree, two concepts must be explained. Therefore, Section 2.2.6 explains the tree pruning and Section 2.2.7 describes the rules applied to stop the splitting. Once the trees are constructed, we want to know if they predict correctly the closing ratio. Thus, Section 2.2.8 presents how to evaluate these CART. This chapter is concluded with a discussion about the advantages and the disadvantage of the CART methodology.

2.2.1. Description

To predict the closing ratio, the statistical method used is the CART model. This technique allows us to find which groups are more likely to buy Meloche Monnex products by partitioning the population into subgroups. A decision tree is a set of questions that splits the data into subgroups depending on the value of the target variable. Consequently, we denote y, the target variable and X, the data set that include the vector of predictors $x = (x_1, ...x_p)^t$, where p is the fixed dimensionality. Therefore, the decision tree begins with all data at the first node, also called the root node. From there, a split criterion based on a particular variable is used to divide the data set into subgroups called the children nodes. For any node t, we suppose that there is a candidate split s^{x_t} of variable x_t. However, some definitions are needed to better understand the decision tree structures.

Definition 2.2.1 (Node). *A **node** t is a partition of the data set. If it is divided into children nodes and it is called a parent node.*

Definition 2.2.2 (Root node). *The **root node** is the complete data set which corresponds to the top node of the tree.*

Definition 2.2.3 (Terminal node). *A **terminal node** is a node with no children nodes.*

Some terminology must be given (see Figure 2.4):

t_L: the left children node,

t_M: the middle children node,

t_R: the right children node,

The proportions of the observations in the different children nodes of the parent node t are given by:

$p_L(t)$: proportion in t that goes into t_L,

$p_M(t)$: proportion in t that goes into t_M,

$p_R(t)$: proportion in t that goes into t_R.

Figure 2.4 shows the CART structure. The root node is at the top of the tree and there is three children nodes with the corresponding probabilities. In this example, because the children nodes are not splitted, they are also terminal nodes.

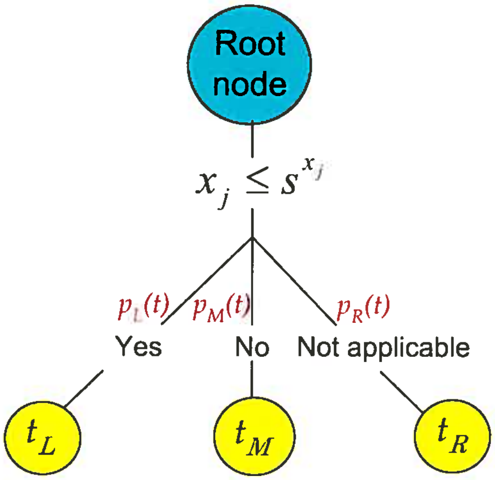


FIG. 2.4. Representation of the CART model. The root node is divided into three children nodes denoted t_L, t_M and t_R.

An important issue in the tree procedure is how to read a tree. The Example 2.1.1 can be used to produce a tree. Figure 2.5 represents a tree that include two split questions. At the beginning, the 10 observations are in the root node and the average closing ratio is 0.38. The first split is produced by the number of auto claims. The 4 observations with less than 2 auto claims are put in the left child node and have an average closing ratio of 0.59. The 4 observations with more than 2 auto claims are in the middle node and have a less important average closing ratio of 0.30. The right child node includes the 2 individuals without an auto policy and with an average closing ratio of 0.13 which is the lowest among the nodes. We note that their closing ratio is only for the residential part. At this point, the middle and the right child nodes are not divided. Consequently, these two nodes are terminal nodes. However, the left node is splitted with the variable "age". Therefore, the observations younger than 40 go in the left node and the others are put in the right node. We can see that this tree allowed us to find a group with a much greater closing ratio than the average population. Indeed, the group with less than 2 auto claims and younger than 40 has an average closing ratio of 0.88.

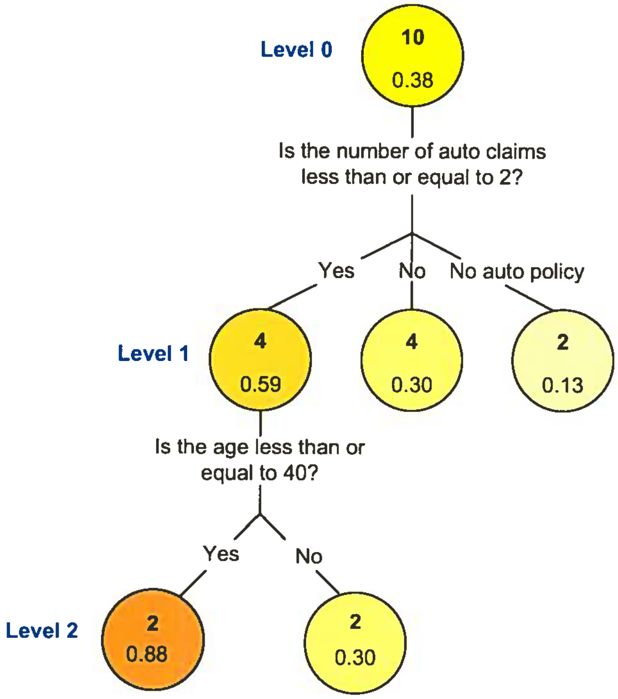


FIG. 2.5. Example 2.1.1: tree model on 10 observations with the variables "age" and "number of claims". Within each node, the first number is the number of observations present and the second corresponds to the average closing ratio. The darker are the nodes, the greater is the average closing ratio within it.

In this chapter, we present the method for CART models for the target variable as a binary variable. Hence, we consider a decision rule (presented in Chapter 3, Section 3.8) that assign 0 for individuals with a small closing ratio, and 1 otherwise. Therefore, we are in a context of a two-categories target variable.

2.2.2. Construction

First of all, the data used to create the model is divided into three groups: the training set (70%) to build a set of models, the test set (30%) to see how the model performs on unseen data.

An important notion in the construction of decision trees is the homogeneity.

Definition 2.2.4 (Homogeneity). *A node is* ***homogeneous*** *when all the observations in it are from the same category. In our context, an homogeneous node will be a node that includes all buyers or all individuals that do not buy any insurance product.*

At the beginning of the tree algorithm, all data is included in the root node. At this point, the objective is to divide this node into children nodes to obtain homogeneous groups in term of closing ratio. Then, a split criterion using a particular variable is used to split the data set. This criterion, also called the splitting rule, must be chosen to perform the best split. At each node the tree algorithm searches through the variables one by one, beginning with x_1 and continuing up to x_p. For each variable it finds the best split. Then it compares the p best single variable splits and selects the best of the best. In the next step, one or more of these regions are split and this process is continued until some stopping rule is applied. We note x_t the variable chosen to split the node t and s^{x_t} is the split value to execute the split.

Definition 2.2.5 (Splitting rule). *A* ***spliting rule*** *is a criterion that divides the data and that is composed with two elements: the variable used to split, and the split-point to achieve the best split.*

Definition 2.2.6 (Stopping rule). *A* ***stopping rule*** *is a splitting rule that makes a node a terminal node.*

The construction of a tree revolves around these elements:

(1) the choice of the best variable to split the data,
(2) the definition of a set of questions Q for each variable,
(3) the selection of the splits by evaluating the goodness of split for any split s^{x_t} of any node t,
(4) the decision when to declare a node terminal or to continue to split,
(5) the assignment of each terminal node to a class.

2.2.3. Set of questions

The set Q of questions generates a set S of splits s^{x_t} of every node t and every variable x_k, $k = 1, ..., p$. These variables can be quantitative or qualitative. The set of questions Q is defined such as:

(1) each split depends on the value of only a single variable,
(2) for each ordered variable x_t, Q includes all questions of the form "Is $x_t \leq s^{x_t}$?" for all s^{x_t} ranging over the domain of x_t,
(3) if x_t is categorical, taking values in $A = (a_1, ..., a_L)$, the questions are of the form "Is $x_t \in$ of a subset of A ?".

2.2.4. Data considerations for the splits

In the literature, decision trees are binary, *i.e.* each parent node produces two children nodes. Hastie *et al.*(2001) suggests that binary splits are better because multiway splits fragment the data too quickly. Therefore, the next level could include insufficient data. However, in our context, the splits can not always be binary. Indeed, the observations in the data set are divided in three categories:

(1) individuals with auto products only,
(2) individuals with residential products only,
(3) individuals with both products.

In Chapter 1, we saw that the variables can be auto variables, residential variables or demographic variables. In the Example 2.1.1, the number of claims is an auto variable and the age is a demographic variable. If the variable used to separate the parent node is an auto variable, the observations without auto products does not have a value for this variable. We thus say that the answer to the question is not applicable for these individuals. Therefore, they are put in the right children node. On the other hand, if the answer to the question is "yes" for an observation, it goes in the left children node and if the answer is "no", it goes in the middle node. However, if the splitting rule is formed with a demographic variable, the split is binary because everyone in the data set has a value for these variables.

As we will see in Chapter 4, most of the time, tertiary splits happen at the beginning of the tree. Indeed, after some splits, the individuals without auto or residential products are isolated. The decision tree produced in this analysis has some binary splits and some tertiary splits. It does not affect the tree quality because the data set is very large and the majority of the splits are binary.

2.2.5. Goodness of split

When the splitting rule is chosen, the objective is to have maximum homogeneity within a node. In other words, we want the child nodes to be as "pure" as possible. Let the split s^{x_t} at each node t that makes immediate descendent nodes as "pure" as possible. A pure node is an homogeneous node, *i.e* a node with all the patterns of the same category. Although, it is more convenient to define the impurity rather than the purity of a node.

Definition 2.2.7 (Goodness of split). *A **goodness of split** is a function $\phi(s^{x_t}, t)$ for any split s^{x_t} of any node t used to evaluate if a split produces partitions different enough of the others.*

Definition 2.2.8 (Impurity). *An* ***impure node*** *is a node that is not homogeneous. The impurity is a measure of goodness of split.*

Let $i(t)$ be the impurity of a node t. We want $i(t)$ to be 0 if all of the patterns that reach the node bear the same category label, and to be large if all the categories are equally represented. The split selected is the split that reduce the node impurity the most. In our case, the objective is to find terminal nodes with the most buyers as possible. Therefore, we want to obtain nodes with the smallest impurity as much as possible. The goodness of the split at node t is thus defined as the decrease in impurity given by:

$$\Delta i(s^{x_t}, t) = \begin{cases} i(t) - p_L(t)i(t_L) - p_M(t)i(t_M) - p_R(t)i(t_R) & \\ & \text{if there is three children nodes,} \\ i(t) - p_L(t)i(t_L) - p_M(t)i(t_M) & \text{if there is two children nodes.} \end{cases}$$

Proposition 2.2.1. *For any node t and split s^{x_t},*

$$\Delta i(s^{x_t}, t) \geq 0.$$

We note $p(c|t)$, the fraction of patterns at node t that are in category c , $c = 1, ..., C$ where C is the number of categories. In our case, $C = 2$ but the general case is presented here. The process of finding good splits is implemented in this way:

(1) define the node proportions $p(c|t)$, $c = 1, ..., C$ so that $\sum_{c=1}^{C} p(c|t) = 1$.

(2) Define a measure $i(t)$ of the impurity of t as a nonnegative function of $p(c|t)$.

(3) The best split s is the split for which the decreasing in impurity $\Delta i(s, t)$ is maximal. Consequently, $\Delta i(s, t)$ must be calculated for all variables and all possible split values.

Suppose we have done some splitting and arrived at a current set of terminal nodes. We denote $\widetilde{T}$, the set of terminal nodes

$$\begin{aligned} I(T) &:= \text{the overall tree impurity} \\ &= \sum_{t \in \widetilde{T}} i(t)p(t). \end{aligned}$$

Selecting the splits that maximize $\Delta i(s, t)$ is equivalent to selecting those that minimize the overall tree impurity (Breiman *et al.*, 1984).

Many different impurity functions can be defined for selecting the best split at each node:

(1) entropy,

(2) Gini impurity (variance impurity for two classes),

(3) missclassification impurity.

All these impurity functions are therefore separately considered in our models. However, according to Breiman *et al.*(1984), the properties of the final tree selected are insensitive to the choice of the impurity function.

2.2.5.1. *The entropy impurity*

The entropy impurity is given by:

$$i(t) = -\sum_{c} p(c|t) \log(p(c|t)), \tag{2.2.1}$$

where $p(c|t)$ is the proportion of observations that are in category c in the node t. If all patterns are of the same category, the impurity is 0; otherwise it is positive, with the largest value occurring when the different classes are equally likely.

2.2.5.2. *The Gini impurity*

The Gini impurity is the expected error rate at node t if the category label is selected randomly from the class distribution present at t. This impurity function is defined as follow:

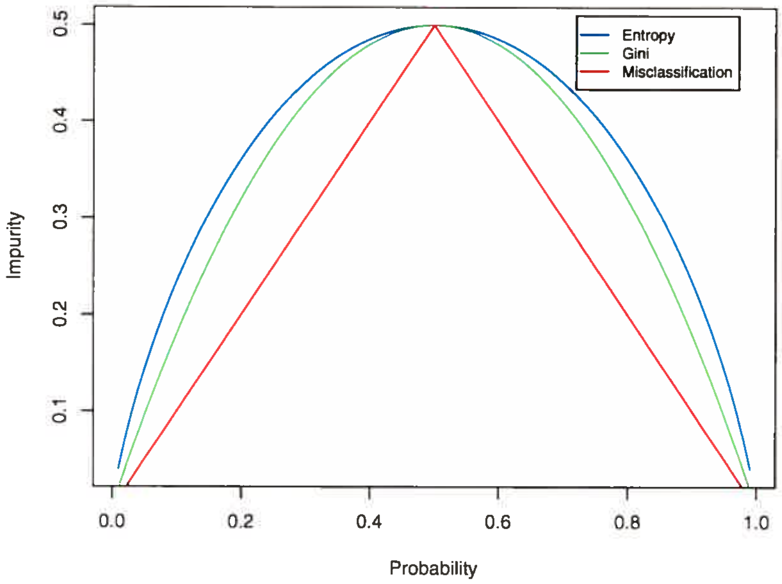


FIG. 2.6. Node impurity measures for two-class classification, as a function of the proportion p in class 2. Entropy has been scaled to pass through (0.5,0.5). (Hastie *et al.*, 2001)

$$\begin{aligned} i(t) &= \sum_{c \neq l} p(c|t)p(l|t) && (2.2.2) \\ &= \sum_{c} p(c|t) \sum_{l \neq c} p(l|t) && (2.2.3) \\ &= \frac{1}{2}[1 - \sum_{c} p^2(c|t)], && (2.2.4) \end{aligned}$$

where $c = 1, ..., C$ represents the observation category. If all observations are in the same category, the impurity will be zero. This measure is the best for a small number of classes and works well for noisy data (Hastie *et al.*, 2001). It is also simple and quick to compute.

In the two-class problem, the index reduces to:

$$i(t) = 2p(1|t)p(2|t).$$

In the two categories case, the Gini impurity is seen as a generalization of the variance impurity (Duda *et al.*, 2001) defined as:

$$i(t) = p(1|t)p(2|t).$$

2.2.5.3. *The misclassification impurity*

An other impurity measure is given by the misclassification impurity. It is defined as:

$$i(t) = 1 - \max_c p(c|t). \tag{2.2.5}$$

It measures the minimum probability that an observation would be misclassified at node t. Among all the impurity measures typically considered, this measure is the most strongly peaked at equal probabilities (Duda *et al.*, 2001). This is represented in Figure 2.6.

2.2.6. Pruning

When the CART model is developed, the tree size must be determined. A very large tree might overfit the data, while a small tree might not capture the important structure. Furthermore, the rules derived from decision trees, especially from large trees are often quite complicated. This is why the trees must be reduced to ease its interpretation.

Definition 2.2.9 (Branch)**.** *(Breiman* et al.*, 1984) A* ***branch*** T_t *of a tree* T *consists of the node* t *and all descendants or children of* t *in* T*.*

Definition 2.2.10 (Pruning)**.** *(Breiman* et al.*, 1984)* ***Pruning*** *a branch* T_t *from a tree* T *consists of deleting from* T *all descendants of* t*.*

Definition 2.2.11 (Horizon effect)**.** ***Horizon effect*** *happens when the splitting is stopped and the model suffers from the lack of sufficient look ahead (beneficial splits in subsequent nodes).*

Therefore, the pruning is executed when the maximum tree is obtained and it is often used because it avoids the horizon effect.

When a tree is pruned, it is grown fully until leaf nodes have minimum impurity as shown in Figure 2.7. Once the big tree has been constructed, the resulting terminal nodes are examined. If the splits are not interesting enough, the nodes are eliminated, because of business reason or because the decrease in impurity is too small. Figure 2.7 is an illustration of this process. It represents the tree of Example 2.1.1 fully developed. The dashed lines show splits that form groups with small average closing ratio. Because the main goal of this project is to identify the buyers, we are not interested in nodes with small closing. Therefore, these nodes could be eliminated and their parent become terminal nodes.

2.2.7. Stopping rules

When a tree is grown, there is a moment when the splitting must stop. If the tree is built until each node corresponds to the lowest impurity, the data have been overfit. Therefore, a stopping rule must be applied at this point. For the choice of the right size tree, many stopping rules could be used:

(1) optimization by minimum number of points,
(2) minimal change in impurity.

2.2.7.1. *Optimization by minimum number of points*

In this method, the splitting is stopped when the number of observations in the node is smaller than a prespecified minimal size. In practice, the minimal size is set to 10% of the learning sample size. This approach is very fast, easy to use and it leads to consistent results.

2.2.7.2. *Minimal change in impurity*

A good way to cut off insignificant nodes is to continue until the change in impurity is too small. If the threshold for the reduction in impurity is noted

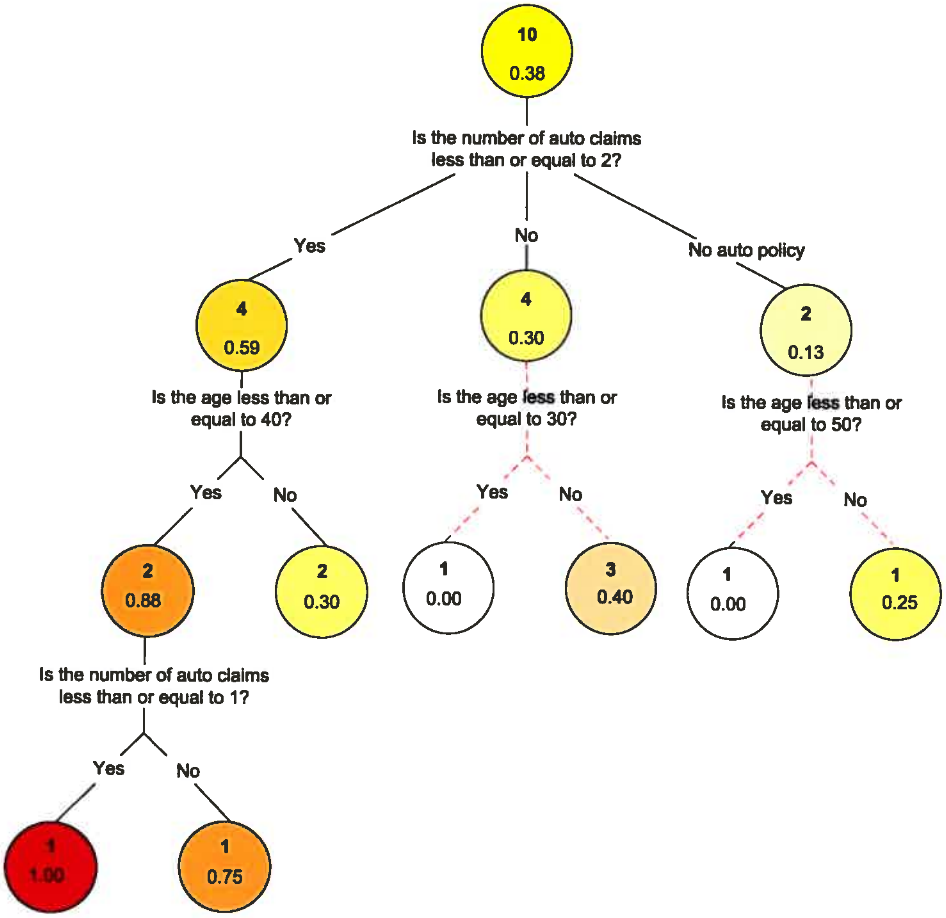

10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Is the age less than or equal to 30?
Is the age less than or equal to 50?
Yes
No
Yes
No
Yes
No
2
0.88
2
0.30
1
0.00
3
0.40
1
0.00
1
0.25
Is the number of auto claims less than or equal to 1?
Yes
No
1
1.00
1
0.75

FIG. 2.7. Example of tree fully grown. Each terminal node is completely pure because all observations in it have the same closing ratio. Therefore, the impurity of each node is null. The dashed connect lines represent splits that could be eliminated because they produce children nodes with small closing ratios. Indeed, the objective is to find which persons have a closing ratio near 1.

$\beta > 0$, the node is declared to be terminal if

$$\max_{s \in S} \Delta i(t) < \beta.$$

However, it is often difficult to know how to set the threshold. Indeed, there is rarely a simple relationship between β and the overall performance (Duda *et al.*, 2001).

2.2.8. Model assessment

In the previous subsections, we developed a model to predict the target variable y. Now, we want to assess this model. Let $\widehat{f}(X)$ be the estimated closing ratio where X represents the set of the predictor variables $x_1, ..., x_p$. This function $\widehat{f}(X)$ has been estimated from the training sample. Therefore, the loss function for measuring errors between y and $\widehat{f}(X)$ is denoted by $L(y, \widehat{f}(X))$. In Hastie *et al.*(2001), they consider two choices for this loss function:

(1) the quadratic loss

$$L(y, \widehat{f}(X)) = (y - \widehat{f}(X))^2,$$

(2) the absolute value loss

$$L(y, \widehat{f}(X)) = |y - \widehat{f}(X)|.$$

The objective is to calculate the test error that is, the expected prediction error over an independent test sample (Hastie *et al.*, 2001):

$$\mathrm{Err} = \mathrm{E}[L(y, \widehat{f}(X))].$$

In Hastie *et al.*(2001), they discuss many methods to estimate this expected value such as:

(1) the bias-variance decomposition,

(2) the optimism of the training error rate,

(3) the estimates of In-sample prediction error,

(4) the cross-validation.

However, as we saw in Subsection 2.2.2, the data set is divided into two samples. Therefore, the training sample allows to construct the model. The test set is used to evaluate the model by calculating:

$$\overline{\mathrm{Err}} = \frac{1}{n}\sum_{j=1}^{n} L(y_j, \widehat{f}(\vec{x_j})),$$

where n is the number of observations, y_j is the closing ratio for observation j, $\vec{x_j} = (x_{1j}, ..., x_{pj})$ and $\widehat{f}(\vec{x_j})$ is the predicted closing ratio for observation j.

Since enough data is available in our context, cross-validation is not needed to estimate the sample error. Indeed, this method uses part of the available data to fit the model and a different part to test it.

2.2.9. Advantages and disadvantages (Breiman *et al.*, 1984)

Decision trees have the potential for being a powerful and flexible classification method. Therefore, it can be applied to any data structure through the appropriate formulation of the set of questions Q. Furthermore, it handles both quantitative and categorical variables. However, the most important advantage of decision trees is their interpretability.

Definition 2.2.12 (Interpretability). *The **interpretability** is when feature space partition is fully described by a single tree. An interpretable tree provides insight and understanding into predictive structure of the data.*

Other advantages of decision trees are interesting:

(1) CART models lead to rapid classification: employing a sequence of typically simple queries,
(2) CART provide a natural way to incorporate prior knowledge from human experts,
(3) nonparametric model because this method does not require specification of any functional form,
(4) does not require variables to be selected in advance. It identifies the most significant variables and eliminate non-significant ones,
(5) CART results are invariant to monotone transformation of its independent variables. If some transformations are applied to the data, the structure of the tree is unchanged. In this case, only splitting value are modified,

(6) CART can easily handle outliers. These observations are isolated in a separate node and they thus don't affect the rest of the model.

(7) On a given level of the CART, there is no restriction for the splitting rule. Therefore, all variables can be used to split at any level.

(8) For all variables, there is no restriction in the split values.

CART models also have some disadvantages:

(1) decision trees may be unstable. They have indeed a high variance. Therefore, insignificant modifications could lead to radical changes. At any node, there may be a number of splits on different variables that give almost the same decrease in impurity. Therefore, the choice between competing splits is difficult and almost random. This choice can lead to a very different interpretation.

(2) CART split only by one variable so they may not catch the correct structure of the data.

In this chapter, two statistical methods were explained. In the first section, cluster analysis was used to classify the individuals into different groups and was described in Section 2.1.1. To understand cluster analysis, a method to calculate the distances was then defined. Therefore, many measures of similarity or dissimilarity between the individuals being clustered were described in Section 2.1.3. The first section continued with the elaboration of two clustering strategies: the bottom-up clustering and the K-means clustering. Consequently, Section 2.1 concluded with the explanation of the method used to find this number.

Section 2.2 elaborated about CART models that are used to predict the closing ratio. In Section 2.2.1, the model structure was described as a tree with a root node at the top and children descendent nodes. Then, the method to construct this model was presented in Sections 2.2.2 and 2.2.3. To evaluate the goodness of the split, the node impurity was calculated. Furthermore, many impurity functions could be used. Therefore, the entropy, the Gini and the misclassification rate were explained in Section 2.2.5. When the tree is formed, an interesting technique is the pruning process. In this case, the tree is fully built and the insignificant nodes are eliminated later. This was presented in Section 2.2.6. An other consideration in the tree construction is to know when the splitting must stop. Thus, some stopping rules were presented in Section 2.2.7. To conclude Chapter 2, the model assessment method was explained as the computation of the prediction error on the test sample.

Chapter 3 has the same objective of Section 2.2. Indeed, it will produce decision trees but using a Bayesian approach instead of the classical approach.

Chapter 3

BAYESIAN APPROACH

In Chapter 2, we explained two useful statistical methods in the business context. Although cluster analysis does not create models, it allows us to better understand the population under consideration. On the other hand, decision trees produce models to predict the target variable, that is the closing ratio in our context. Thus, decision trees allow us to have a better understanding of the buyers characteristics.

However, in the business situation, the trees grown in Chapter 2 with the classical approach has a main disadvantage. Indeed, each variable has a probability to be selected without regard to his business importance. In the opposite of this approach, the Bayesian approach weights each variable depending of his importance. Therefore, it is easier to obtain trees that are easy to explain and understand. In our context, interpretable trees are essential if we want them to be used on the actual population.

In this chapter, the Bayesian approach is used to add flexibility in the model. This approach uses the *a priori* information to construct statistical models. Therefore, we introduce prior specification on all the unknowns. This allows us to minimize undesirable model characteristics, such as tree complexity, or express a preference for certain predictor variables. In the construction of decision trees, the Bayesian approach is called BART, meaning "Bayesian regression tree". In that case, it is used to increase the interpretability of the tree. By modeling the unknown parameters of the sampling distribution through a probability structure, the Bayesian approach allows us to give more weight to some variables useful to the subject matter specialist.

This chapter is concerned with one main issue: to find trees with high probability, more precisely with high posterior probability. Therefore, many trees must be constructed in order to find the best model. We first start with the presentation of some aspects of the Bayesian paradigm. We continue in Section 3.2 by explaining how the tree is structured. Furthermore, the Example 2.1.1 is used to illustrate these notions. In order to calculate the posterior probability of a tree, the specification of two elements are essential: the distribution of the target variable y and the tree prior $\mathbb{P}(T)$. To find the y distribution, the predictor priors are given in Section 3.3 and the distribution of y is presented in Section 3.4. The specification of the tree prior is done in Section 3.5 by presenting three priors: the prior for the choice of the split variable, the prior for the choice of the split value and finally, the prior for the number of terminal nodes. The tree posterior is thus presented in Section 3.6 with the variable and split posteriors. The algorithm applied to construct the trees is then described in Section 3.7. Once the trees are created, the categories must be assigned to terminal nodes. This is presented in Section 3.8. This chapter concludes by a discussion on many criterions to evaluate the trees such as the trees posterior probability, the likelihoods $\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T})$ and finally, the misclassification rates. Therefore, the best tree can be chosen.

3.1. Bayesian theory

Consider the model represented by $y \sim f(y|\beta)$ where y is the vector of the observations and β is the vector of parameters. The parameters β could be modeled through a probability distribution $\pi(\beta)$, called prior distribution.

Definition 3.1.1 (Bayesian statistical model). *A Bayesian statistical model is made of a parametric statistical model, $f(y|\beta)$, and a prior distribution on the parameters, $\pi(\beta)$ (see Robert, 2001).*

The inference is based on $\pi(\beta|y)$, the distribution of β conditional on y, called posterior distribution and defined by

$$\begin{aligned}\pi(\beta|y) &= \frac{f(y|\beta)\pi(\beta)}{\int f(y|\beta)\pi(\beta)d\beta} \\ &= \frac{f(y|\beta)\pi(\beta)}{m(y)},\end{aligned}$$

where $m(y)$ represents the marginal density of y defined below.

Definition 3.1.2 (Marginal distribution). *The marginal distribution of y is the density function of y alone, integrating the information about β. This distribution function is:*

$$m(y) = \int f(y|\beta)\pi(\beta)d\beta.$$

3.2. BART STRUCTURE

A BART model is a Bayesian statistical model that describes the conditional distribution of the target variable y given a vector of predictors $x = (x_1, x_2, ..., x_p)^t$ where p is the dimension of the vector. A Bayesian tree model has two components: the tree T with b terminal nodes and a vector of parameters $\Theta = (\theta_1, \theta_2, ..., \theta_b)^t$. The parameter value θ_i is associated to the i^{th} terminal node. If x lies in the region corresponding to the i^{th} terminal node, then the variable $\vec{y}$ given $\vec{x}$ has distribution $f(\vec{y}|\theta_i)$, where f represents a parametric family indexed by θ_i. We set $\theta_i = (x_i, s^{x_i})$, where x_i is the chosen variable for node i and s^{x_i} is the chosen split value for node i. Furthermore, we denote y_{ij} as the j^{th} observation of y in the i^{th} terminal node, $j = 1, ..., n_i$ and $i = 1, ..., b$ where n_i is the number of observations in node i. The b regions corresponding to the b terminal nodes are disjoint so that the tree separates the data set by assigning each observation to one of the b terminal nodes. Consequently, if we denote the total number of observations in the data set by n, we have that $\sum_{i=1}^{b} n_i = n$. To illustrate these concepts, a BART model is applied on Example 2.1.1 and the tree is represented in Figure 3.1. In this example, the data consists of 10 observations and two predictor variables: the individual's age and his number of auto claims (see table 2.1). Therefore, $x = (x_1, x_2)$. The Figure 3.1 shows a tree with $b = 4$ terminal nodes and $n_1 = 2$, $n_2 = 2$, $n_3 = 4$ and $n_4 = 2$. Consequently, we see that each observation is assigned to one particular terminal node.

As we explained in Chapter 2, each tree node forms a partition of the data set.

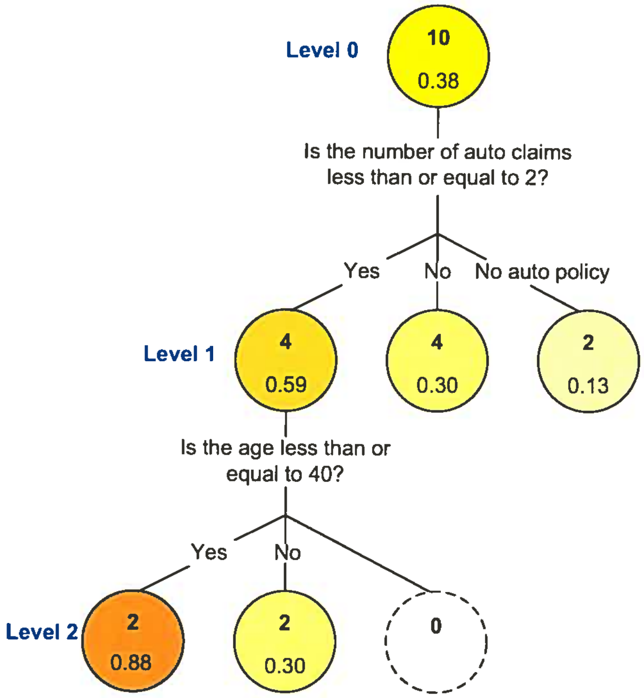

Level 0
10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
Level 1
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Yes
No
Level 2
2
0.88
2
0.30
0

FIG. 3.1. Example 2.1.1: tree model on 10 observations with the variables "age" and "number of claims". Within each node, the first number is the number of observations present and the second corresponds to the average closing ratio. Each question produces three splits even though it may produce an empty node (the dashed node).

Therefore, in order to identify each of them, we define a label on each node. We note the position of each node by t. The root node, which is always in the model, is chosen to be the first split node and its position is labeled as position 0, so that $t = 0$. Any descendant splitting node's position, t, is uniquely defined given its parents's position, t^{parent}. As we saw in Chapter 2, a parent node can produce either two or three children nodes. In order to produce a model without too much complexity, we always produce three descendant splits. If the split produces only two subsets, we set the third descendant node to be null. This allows us to obtain a consistent terminology. Furthermore, the null node does not count for a terminal node because it exists just for computationally matter. For example, in Figure 3.2, we see that the node with label 6 is null. However, even if this node is null, it is interesting to identify it by its label. Therefore, we define the labels for the three children nodes as follow. For a node t, given the parents's position, its label is given by:

$$t = \begin{cases} 3t^{\text{parent}} + 1 & \text{if the node contains the data points} \\ & \text{for which the question at the parent node is true,} \\ 3t^{\text{parent}} + 2 & \text{if the answer to the question is no,} \\ 3t^{\text{parent}} + 3 & \text{if the question is not applicable.} \end{cases}$$

To better understand the labels, the Figure 3.2 shows how the tree is structured. Indeed the children label depends on its parent's label and on the split question answer. Figure 3.1 can also be used. The labels allow us to identify each particular subset. For example, the terminal node with individuals having less than 2 auto claims and older than 40 years old is identified as the node labeled 5. Furthermore, this terminology allows to identify the node with the highest average closing ratio. Consequently, we see that node 4 has an average closing ratio of 0.88 that is the greatest average among the tree.

Another consideration is the tree levels L. At the top of the tree, the root level is $L = 0$ and after the first split, $L = 1$. Using the previous notation, the labels can easily be defined on a same level. On level L, there are exactly 3^L nodes. The first node label on a given level is the total number of nodes on the previous levels. This is because the root node is labeled 0. Therefore, the first node on a level L is given by:

$$l_t = \sum_{m=0}^{L-1} 3^m = 3^0 + ... + 3^{L-1}$$

$$\Longrightarrow l_t = \frac{3^L - 1}{2}.$$

The last position on a level L will be $\frac{3^L-1}{2} + \frac{2*3^L-2}{2}$. The positions of the nodes on a level L are therefore $l_t = \frac{3^L-1}{2}, ..., \frac{3^{L+1}-3}{2}$. For example, Figure 3.2 shows that the labels on level 2 are from 4 to 12. This figure also shows the node labels and the tree levels. If the parent node has a label of 3, the children labels will be 10, 11 and 12. In the figure, we can also see that some nodes can be empty. Therefore, the nodes 6 and 9 do not have any observation.

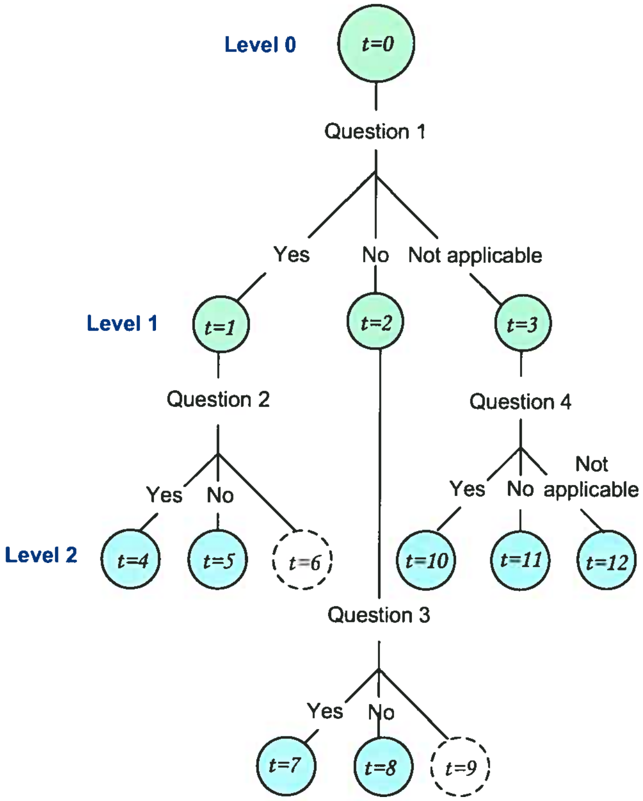


FIG. 3.2. BART with two levels and 7 terminal nodes. The labels are shown within the nodes. The blue nodes are terminal nodes, the green are nonterminal and the white dashed represents a null node. The tree levels are also represented.

3.3. Distributions of the predictors

In order to develop a model for the distribution of y, we need to know the distributions of the predictor variables. Thus, the likelihood is a function of the distribution of these predictors. Therefore, we have to specify each variable density. Because the variables used are either continuous or discrete, this distribution depends on the nature of the variable.

For the continuous variables, the distribution is determined by plotting the histogram of each variable. Therefore, three continuous distributions are considered:

(1) If the histogram shows a "bell-shaped curve" which is symmetrical about the mean as Figure 3.3, the variable $x_t \sim N_p(\theta, \sigma^2)$. In this case, the density of x_t for $\sigma > 0$ is given by:

$$\pi(x_t|\theta, \sigma) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{(x_t-\theta)^2}{2\sigma^2}}.$$

(2) If the histogram shows a non symmetric curve with a long right tail as Figure 3.4, x_t follows a gamma density with parameters α and β such as $x_t \sim G(\alpha, \beta)$. The distribution of x_t is given by:

$$\pi(x_t|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x_t^{\alpha-1} e^{-\beta x_t} \mathbb{I}_{[0,+\infty]}(x_t),$$

where $\alpha, \beta > 0$.

(3) If the histogram is bimodal like Figure 3.5, the distribution of x_t is therefore a mixture of two Gamma such as $x_{t1} \sim G(\alpha_1, \beta_1)$ and $x_{t2} \sim G(\alpha_2, \beta_2)$. The x_t variable is defined as:

$$x_t = \begin{cases} x_{t1} & \text{with probability } 1-p, \\ x_{t2} & \text{with probability } p, \end{cases}$$

where p is the proportion of observations that follow $G(\alpha_2, \beta_2)$. The density of x_t is then given by:

$$\pi(x_t|p) = (1-p)\pi(x_{t1}|p) + p\pi(x_{t2}|p),$$

where $0 < p < 1$, $\alpha_1 > 0$, $\beta_1 > 0$, $\alpha_2 > 0$, $\beta_2 > 0$.

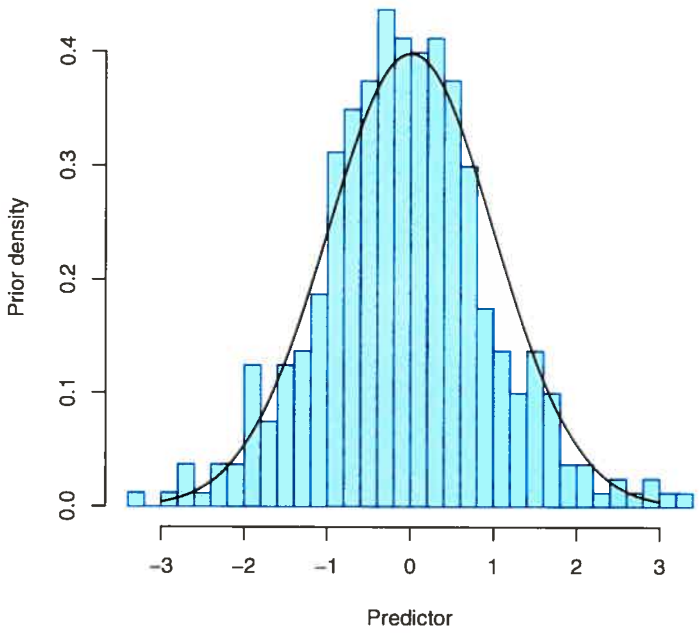


FIG. 3.3. Histogram for a normal density with $\theta = 0$ and $\sigma = 1$.

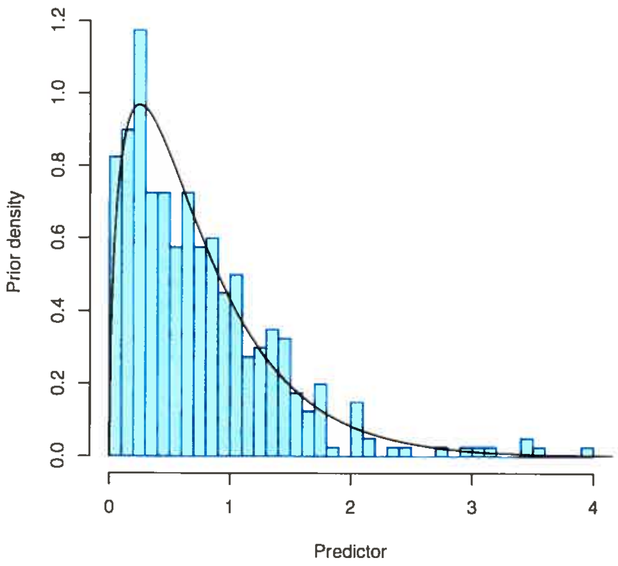


FIG. 3.4. Histogram for a gamma density with $\alpha = 1.5$ and $\beta = 2$.

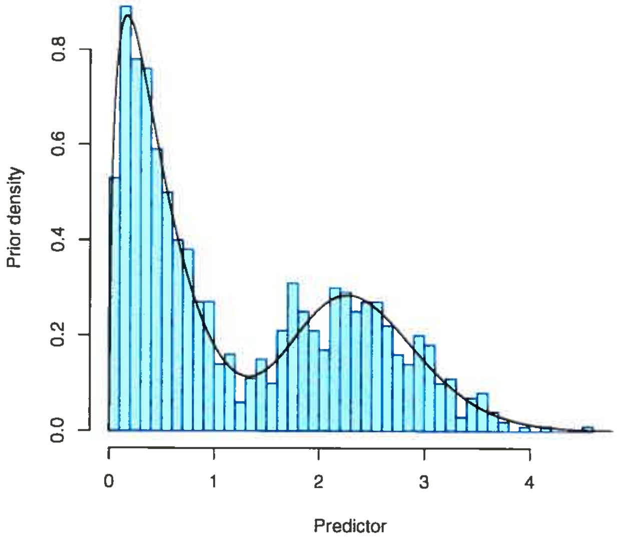


FIG. 3.5. Histogram for a mixture of two gamma densities with $\alpha_1 = 1.5$, $\beta_1 = 3$, $\alpha_2 = 17$, $\beta_2 = 7$ and $p = 0.4$.

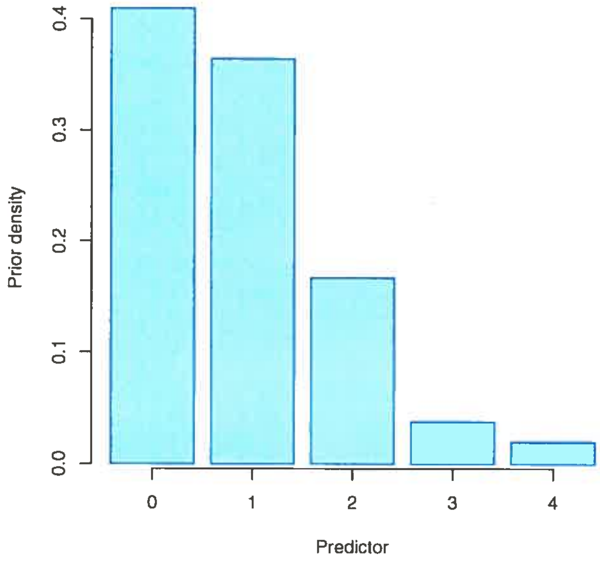


FIG. 3.6. Bar chart of a Poisson density with parameter $\lambda = 1$.

For the discrete variables, two possibilities are considered. Indeed, the discrete variables can be either binary variables or quantitative discrete variables. Hence, the variable density depends on its type. To model the binary variables, a Bernouilli distribution is appropriate. In this case, $x_t \sim Ber(p)$ and its density is given by:

$$\pi(x_t|p) = p^{x_t}(1-p)^{1-x_t}\mathbb{I}_{[0,1]}(x_t),$$

where $0 < p < 1$. The quantitative discrete variables are modeled as Poisson variables. In Example 2.1.1, the variable "number of auto claims" should be modeled as this distribution. Indeed, the Poisson variables are the number of events occurring in a fixed period of time with a known average rate. Furthermore these variables are independent of the time since the last event. If $x_t \sim P(\lambda)$, the distribution of x_t for $\lambda > 0$ is given by:

$$\pi(x_t|\lambda) = e^{-\lambda}\frac{\lambda^{x_t}}{x_t!}\mathbb{I}_{\mathbb{N}}(x_t).$$

This density is shown in Figure 3.6.

To estimate the hyperparameters, the maximum likelihood method is used. The gamma parameters are found using an iteration scheme. For the mixture of the two gamma densities, the EM algorithm described in Chapter 1 is used to calculate each distribution weight and also each hyperparameter value.

3.4. Distribution of the variable of interest y

The objective of this project is to create a model to predict the response variable that is the closing ratio. Furthermore, one individual closing ratio depends on his number of purchases among his number of quotes. Hence, we can consider the closing ratio distribution as a binomial distribution depending on the number of quotes and sales. This distribution depends also on the variable x_t used to predict the closing ratio within the node t. Therefore, we consider the link function of a predictor as the associated binomial probability.

Definition 3.4.1 (link function). *The link function is a function that provides the relationship between the predictor and the response variable distribution function.*

An important property of the link function is that it is used to model the predictor variables when the dependent variable is assumed to be nonlinearly related to the predictors (McCullagh and Nelder, 1989). Various link functions are commonly used, depending on the assumed distribution of the dependant variable y. However, it is important to match the domain of the link function to the range of the distribution function's mean. In our case, the link function is a probability function so the domain must be between 0 and 1. For the variable x_t, the link function used in this project is given by:

$$\begin{aligned} p(s^{x_t}, x_t) &= \mathbb{P}(x_t \leq s^{x_t}) \\ &= \int_{\min(x_t)}^{s^{x_t}} \pi(x_t) dx_t. \end{aligned}$$

The y distribution depends on each observation distribution functions within all terminal nodes. Indeed, the union of all terminal nodes at a given level includes all the observations of the data set. Therefore, the y distribution is obtained by multiplying the distribution function of each observation in a given terminal node. For BART models, it is assumed that the response variable values within a terminal node are i.i.d. Furthermore, these y values across terminal nodes are independent (Chipman and McCulloch, 1998). We denote $y_{ij} = (q_{ij}, v_{ij})$ where q_{ij} and v_{ij} are respectively the number of quotes and the number of sales for individual j in node i. The distribution of y_{ij} is given by:

$$f(y_{ij}|x_i, s^{x_i}) = \binom{v_{ij}}{q_{ij}} p(s^{x_i}, x_i)^{v_{ij}} [1 - p(s^{x_i}, x_i)]^{q_{ij} - v_{ij}}, \tag{3.4.1}$$

where x_i corresponds to the variable used to obtain node i and s^{x_i} is its split value. Therefore, the y distribution within a node i is simply the product of equation (3.4.1) for all terminal nodes.

$$f(\vec{y_i}|x_i, s^{x_i}) = \prod_{j=1}^{n_i} f(y_{ij}|x_i, s^{x_i}), \tag{3.4.2}$$

where n_i is the number of observations in node i.

3.5. Specification of the tree prior

The tree prior is specified by a tree-generating stochastic process. Each realization of such a process can simply be considered as a random draw from this prior. To draw from the prior, we start with the tree consisting of a single root node. The tree then grows by randomly splitting terminal nodes by assigning them splitting rules and children nodes.

Therefore, the growing process is determined by the specification of three functions. First, we need the prior of the variable chosen to split the data and second, the prior of the split value given this variable. An important consideration is also the size of the tree, *i.e.* the choice of the number of terminal nodes. Thus, the tree prior probability consists of the product of the following priors: the prior for the choice of the variable, the splitting value prior and finally, the prior for the number of terminal nodes. These three priors are multiplied together to obtain the tree prior denoted by $\mathbb{P}(T)$. If a tree has b terminal nodes, then there is b ways to reach the final level of the tree. For each way, a prior probability is calculated. Therefore, the tree prior is obtained as follow:

$$\mathbb{P}(\vec{T}) = \left\{ \prod_{i=1}^{b} \mathbb{P}(\text{choosing variable } x_i)\mathbb{P}(\text{choosing split value } s^{x_i}|x_i) \right\} \\ \times \mathbb{P}(b \text{ terminal nodes}), \tag{3.5.1}$$

where x_i is the variable that splits the data to obtain the terminal node i.

The next subsections will explain each of the following prior:

(1) $\mathbb{P}(\text{choosing variable } x_i)$ that correspond to the priors on variables,

(2) $\mathbb{P}(\text{choosing split value } s^{x_i}|x_i)$, that are the prior on split value,

(3) $\mathbb{P}(b \text{ terminal nodes})$ that is the prior on the number of terminal nodes.

3.5.1. Prior on variables

When the tree is grown, the first step is the selection of the variable used to split the data set. The prior on variable is therefore an important consideration and many priors are possible. A popular choice for this prior is the distribution obtained by choosing x_t uniformly from available predictors. In this case, the prior is given by:

$$\pi(x_t) = \frac{1}{p},$$

where p is the number of predictor variables. This choice represents the prior information that at each node, available predictors are equally likely to be effective. This prior is very simple and also invariant to monotone transformations on the quantitative predictors.

However, in the business situation, some variables are thought to be more important than others. Therefore, it could be better to consider a prior that takes into consideration the weight of each variable. In order to calculate a weight function, we note w_k the variable x_k weight so that:

$$\sum_{k=1}^{p} w_k = 1.$$

The weights are determined as follow. Giving its importance, each variable x_k gets a score a_k from 0 to 10. The most important variable has a score of 10 and the less important gets a score of 1. A score of 0 means that the variable is simply eliminated from the model. For example, if the variable is constant for all observations, its score will be null because it does not add any information to the model. The variable weight is calculated as follow:

$$w_k = \frac{a_k}{\sum_{k=1}^{p} a_k}. \tag{3.5.2}$$

and the prior of variable x_k is given by:

$$\pi(x_k) = w_k. \tag{3.5.3}$$

3.5.2. Prior on splitting values

Once the split variable is chosen, the split value must be found. We denote $\pi_2(s^{x_t}|x_t)$, the prior distribution on splitting value x_t. Because the choice of a split value depends on the predictor variable, the distribution of the split value is a distribution on the set of the predictors. Chipman and McCulloch (1998) considered only priors for which the overall set of possible split values is finite. In this case, $\pi_2(s^{x_t}|x_t)$ is a discrete distribution. However, in this project, we consider this distribution as a continuous function on the values of the variable. Thus, even though we consider the split variable as being a continuous random variable, it is sufficient, in practice, to only consider the observed values as potential split. For example, if we observed $s_{(1)} < s_{(2)} < ... < s_{(k)}$, the split of the data set will be the same if we choose any values between $s_{(i)}$ and $s_{(i+1)}$.

In Chipman and McCulloch (1998), the uniform specification for this prior is presented. Thus, the prior on a split value s^{x_t} given the chosen variable x_t is given by:

$$\pi_2(s^{x_t}|x_t) = \frac{1}{N^{x_t}},$$

where N^{x_t} is the number of x_t possible values. This method assigns lower probability to splitting values based on a variable with more potential split values. In Chipman and McCulloch (2002), they discuss the conflict between the prior and the likelihood information from the data. If the prior is too concentrated around its prior mean, it may be too informative. This can overwhelm the information in the data corresponding to a terminal node. However, if the prior is too spread out, the posterior distribution will also be spread out, particularly for trees with many terminal nodes. A spread out prior is called a diffuse prior. It assigns small probability to each individual split value for variables with many possible values. Furthermore, the probability on each split value is much more important for variables with fewer possible values.

We prefer to use the distribution on split value given the variable chosen as an uniform on the range of the variable. The prior of a split value s^{x_t} given the chosen variable x_t is therefore:

$$\pi_2(s^{x_t}|x_t) = \frac{1}{x_t^{max} - x_t^{min}}, \tag{3.5.4}$$

where x_t^{max} and x_t^{min} are respectively the maximum and the minimum of the split variable.

3.5.3. Prior for the number of terminal nodes

An important consideration in the construction of tree is its size. A good size tree must not be too small or too big. In Chipman and McCulloch (1998), they first consider the prior for the number of terminal node as a constant as follow:

$$\mathbb{P}(\eta = b) = \alpha,$$

where $\alpha > 0$, $b > 0$ and η represents the number of terminal nodes. Under this prior, all trees with b terminal nodes have the same probability regardless of their depth and shape. The shape is defined as the number of consecutive splits above the terminal node. To consider the depth of the tree in this probability, another form is proposed:

$$\mathbb{P}(\eta = b) = \alpha(1 + l_t)^{-\beta},$$

where $\beta \geq 0$ and l_t is the depth of the node t that is the level of the node t. This prior is a decreasing function of the depth. Therefore, deeper nodes will be less likely to split.

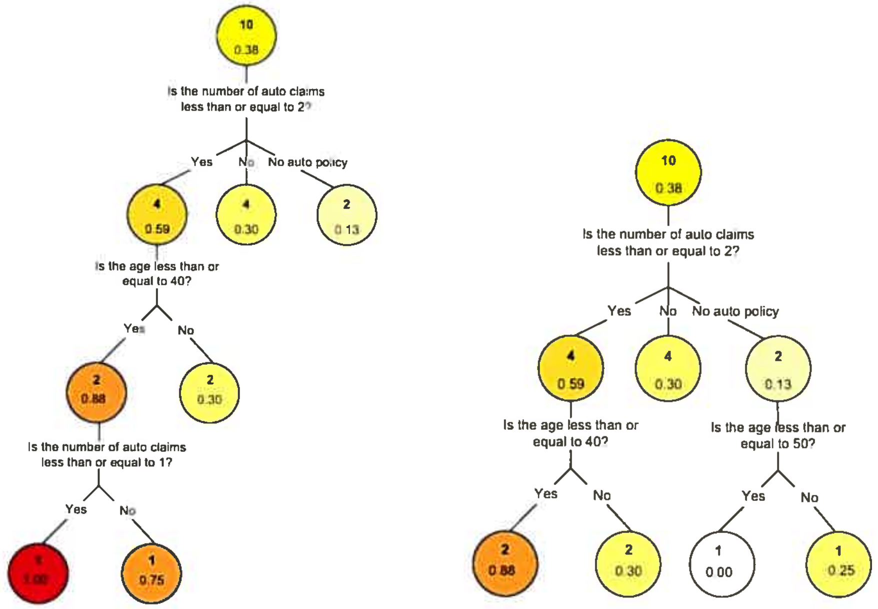
Figure 3.7 shows two different trees with the same number of terminal nodes. Although they have both five terminal nodes, their shapes are very different. In Figure 3.7 (a), the tree is not equilibrated because it is developed only on the left side. On this side, there are three consecutive splits before the terminal node. Figure 3.7 (b) illustrates a tree with the left and right sides equally splitted. This tree is therefore more equilibrated. Indeed, there are two consecutive splits before the last level on the right and left sides.

Instead of using one of these two priors, we chose $\mathbb{P}(\eta = b)$ as a truncated poisson (Denison and Mallick, 2000). In this distribution, the domain of random variable η is restricted to be greater than zero so the minimum number of terminal node is one. The prior for the number of terminal nodes is given by:

$$\mathbb{P}(\eta = b) = \frac{\lambda^b}{b!(e^\lambda - 1)} \qquad \text{for } b = 1, \ldots \tag{3.5.5}$$

When the tree is grown, we need to specify λ to compute this probability. Therefore, we must specify the average number of terminal nodes in the tree. If this number is b, we find λ for which $b = E(\eta)$ where:

$$\begin{aligned} E(\eta) &= \sum_{m=1}^{\infty} \frac{m\lambda^m}{m!(e^\lambda - 1)} \\ &= \frac{1}{(e^\lambda - 1)} \sum_{m=1}^{\infty} \frac{\lambda^m}{(m-1)!} \\ &= \frac{\lambda}{(e^\lambda - 1)} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} \\ &= \frac{\lambda e^\lambda}{(e^\lambda - 1)}. \end{aligned}$$


10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Yes
No
2
0.88
2
0.30
Is the number of auto claims less than or equal to 1?
Yes
No
1
1.00
1
0.75
10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Is the age less than or equal to 50?
Yes
No
Yes
No
2
0.88
2
0.30
1
0.00
1
0.25

(a)

(b)

FIG. 3.7. Example of two trees with the same number of terminal nodes but with different shapes: (a) This tree is not equilibrated because only the left side is developed. On the left side, there is three consecutive splits before the last level while the other sides include only one split before the terminal nodes. Furthermore, there is a total of five terminal nodes. (b) This tree is more equilibrated. Indeed the left and right sides are developed. On the left and right sides, there is two consecutive splits before the terminal node while the middle node has a single split before the terminal nodes. Furthermore, there is a total of five terminal nodes.

So we can now write the tree prior equation (3.5.1):

$$\begin{aligned} \mathbb{P}(\vec{T}) &= \left\{ \prod_{i=1}^{b} w_i \pi_2(s^{x_i}|x_i) \right\} \mathbb{P}(\eta = b) \\ &= \left\{ \prod_{i=1}^{b} \frac{w_i}{x_i^{max} - x_i^{min}} \right\} \frac{\lambda^b}{b!(e^{\lambda} - 1)}. \end{aligned} \tag{3.5.6}$$

3.6. Posterior distributions

Now that all the elements to define the tree posterior distribution are known, we begin this section with its specification.

3.6.1. Posterior distribution on tree

The tree posterior $\mathbb{P}(\vec{T}|\vec{x},\vec{s},\vec{y})$ can be defined as:

$$\mathbb{P}(\vec{T}|\vec{x},\vec{s},\vec{y}) \propto \mathbb{P}(\vec{y}|\vec{x},\vec{s},\vec{T})\mathbb{P}(\vec{T}), \tag{3.6.1}$$

where:

$$\begin{aligned}\mathbb{P}(\vec{y}|\vec{x},\vec{s},\vec{T}) &= \prod_{i=1}^{b} f(y_i|\theta_i) \\ &= \prod_{i=1}^{b}\prod_{j=1}^{n_i} f(y_{ij}|\theta_i).\end{aligned}$$

Using equation (3.4.2), we obtain:

$$\begin{aligned}\mathbb{P}(\vec{y}|\vec{x},\vec{s},\vec{T}) &= \prod_{i=1}^{b}\prod_{j=1}^{n_i} f(y_{ij}|x_i, s^{x_i}) \\ &= \prod_{i=1}^{b}\prod_{j=1}^{n_i} \binom{v_{ij}}{q_{ij}} p(s^{x_i}, x_i)^{v_{ij}}[1 - p(s^{x_i}, x_i)]^{q_{ij}-v_{ij}}.\end{aligned} \tag{3.6.2}$$

Using equation(3.5.6), the tree posterior distribution is given by:

$$\begin{aligned}\mathbb{P}(\vec{T}|\vec{x},\vec{s},\vec{y}) \propto &\left\{\prod_{i=1}^{b} \frac{w_i}{x_i^{max} - x_i^{min}}\right\} \frac{\lambda^b}{b!(1-e^{-\lambda})} \\ &\times \prod_{i=1}^{b}\prod_{j=1}^{n_i} \binom{v_{ij}}{q_{ij}} p(s^{x_i}, x_i)^{v_{ij}}[1 - p(s^{x_i}, x_i)]^{q_{ij}-v_{ij}}.\end{aligned} \tag{3.6.3}$$

3.6.2. Posterior distribution for the choice of the split variable

The first step in the tree construction is the choice of the split variable. This variable must be the one that produce the most different subsets in term of closing ratio. Therefore, the variable chosen is the variable with the highest posterior probability. This conditional probability is calculated by taking into consideration the variable prior and the marginal distribution of y given this variable. We also consider that the variable x_t is the predictor used to split the node t_i that include n_i observations. For variable x_t, its posterior probability is given by:

$$q(x_t|\vec{y}) = \frac{m_2(\vec{y}|x_t)\pi(x_t)}{\sum_{k=1}^{p} m_2(\vec{y}|x_t)\pi(x_t)}, \tag{3.6.4}$$

where $\pi(x_t) = w_t$ is the weight of the variable x_t so that a larger mass is applied on the most important variables and $m_2(\vec{y}|x_t)$ represents the marginal distribution of y given the split variable. We can see that the weight variable has a direct impact on the choice of the split variable. To find the marginal of y given the split variable x_t, all possible splits for this variable must be take into account. The marginal of y given the split variable x_t is:

$$\begin{aligned} m_2(\vec{y}|x_t) &= \int_s f(\vec{y}|x_t, s^{x_t})\pi_2(s^{x_t}|x_t)ds \\ &= \int_s \left[\prod_{j=1}^{n_t} f(\vec{y_t}|x_t, s^{x_t})\right] \frac{1}{x_t^{max} - x_t^{min}} ds \\ &= \frac{1}{x_t^{max} - x_t^{min}} \\ &\quad \times \int_s \prod_{j=1}^{n_t} \binom{v_{tj}}{q_{tj}} p(s^{x_t}, x_t)^{v_{tj}} [1 - p(s^{x_t}, x_t)]^{q_{tj} - v_{tj}} ds. \end{aligned} \tag{3.6.5}$$

The variable chosen is given by:

$$x_t = \underset{x_h \in (x_1, \ldots, x_p)}{\text{argmax}} \; q(x_h|y). \tag{3.6.6}$$

3.6.3. A posterior density for the choice of the split value

We need now to determine the best split given the choice of the best variable to split. As for the choice of the split variable, the choice of the split value is based on the split having the highest posterior probability. Therefore, all possible split values for the variable x_t are considered. To define a range for these splits, the variables are plotted given the response variable. We denote S^{x_t}, the set of the possible splits for the variable x_t. In order to calculate the posterior probability, the equations (3.5.4), (3.6.2) and (3.6.5) are needed. At node t, when the split variable is x_t, the distribution of the split $s_m^{x_t}$ is therefore given by:

$$\begin{aligned}\pi_2(s_m^{x_t}|x_t, y) &= \frac{f(\vec{y}|x_t, s_m^{x_t})\pi_2(s_m^{x_t}|x_t)}{m_2(\vec{y}|x_t)} \\ &= \frac{f(\vec{y}|x_t, s_m^{x_t})\pi_2(s_m^{x_t}|x_t)}{\int_{S^{x_t}} \pi_2(s^{x_t}|x_t) f(\vec{y}|x_t, s^{x_t}) ds^{x_t}} \\ &= \frac{\prod_{j=1}^{n_t} \binom{v_{tj}}{q_{tj}} p(s^{x_t}, x_t)^{v_{tj}}[1 - p(s^{x_t}, x_t)]^{q_{tj}-v_{tj}}}{\int_{S^{x_t}} \prod_{j=1}^{n_t} \binom{v_{tj}}{q_{tj}} p(s^{x_t}, x_t)^{v_{tj}}[1 - p(s^{x_t}, x_{it})]^{q_{tj}-v_{tj}} ds^{x_t}} .\end{aligned}$$

where n_t is the number of observations in the node t, v_{tj} is the number of purchases made by the observation j in the node t and q_{tj} is the number of quotes made by the observation j in the node t.

The better split value for x_t is the split for which:

$$s_l = \underset{s_m^{x_t} \in S^{x_t}}{\text{argmax}} \prod_{j=1}^{n_t} \binom{v_{tj}}{q_{tj}} p(s^{x_t}, x_t)^{v_{tj}}[1 - p(s^{x_t}, x_t)]^{q_{tj}-v_{tj}}. \qquad (3.6.7)$$

3.7. Construction of the trees

In this section, we present the algorithm developed to construct the trees. At first, it starts by splitting the top of the tree using variable x_0 and splitting value $s_m^{x_0}$ where $(x_0, s_m^{x_0})$ represents the best splitting value. It continues by randomly choosing among three independent steps (Denison and Mallick, 2000): the growing, the pruning and the decision to stay. On node t, each of these moves has the respective probabilities b_t, d_t and r_t such as:

$$b_t + d_t + r_t = 1.$$

Furthermore, the different move probabilities are affected by the number of nodes produced at each split. In the growing step, the split can produce either two or three children nodes. On the other hand, in the pruning step, it can delete one or two nodes.

To present the steps, we suppose that we are in node t and that the tree has b terminal nodes at this point. These steps are described in the next sections.

3.7.1. Growing

When a node is grown, it is divided into children nodes by assigning it a splitting rule from the prior. The probability of growing the node t is given by:

$$\mathbb{P}_{\text{GROW}}(t) = b_t = \begin{cases} 2\tau\min\left\{1, \frac{\mathbb{P}(\eta=b+1)}{\mathbb{P}(\eta=b)}\right\} & \text{if there is two children nodes,} \\ 2\tau\min\left\{1, \frac{\mathbb{P}(\eta=b+2)}{\mathbb{P}(\eta=b)}\right\} & \text{if there is three children nodes.} \end{cases}$$

where $\mathbb{P}(\eta = b+1)$ is the probability of having $b+1$ terminal nodes and τ is a constant to be defined later. Therefore, by the equation (3.5.5), the ratio can be written:

$$\frac{\mathbb{P}(\eta = b+1)}{\mathbb{P}(\eta = b)} = \frac{\lambda}{b+1}.$$

It represents the probability of passing from b terminal nodes to $b+1$ terminal nodes. When the split produces three children nodes, the number of terminal nodes increases by two. The ratio can also be simplified by:

$$\frac{\mathbb{P}(\eta = b+2)}{\mathbb{P}(\eta = b)} = \frac{\lambda^2}{(b+1)(b+2)}.$$

3.7.2. Pruning

Pruning means to turn a parent node into a terminal node by collapsing the nodes below it. The probability is:

$$\mathbb{P}_{\text{PRUNE}}(t) = d_t = \begin{cases} \tau\min\left\{1, \frac{\mathbb{P}(\eta=b)}{\mathbb{P}(\eta=b+1)}\right\} & \text{if there is two children nodes,} \\ \tau\min\left\{1, \frac{\mathbb{P}(\eta=b)}{\mathbb{P}(\eta=b+2)}\right\} & \text{if there is three children nodes.} \end{cases}$$

This move is the inverse of the growing move. Indeed, if a node is pruned, the number of terminal nodes decreases. If there is two children nodes, these two are collapsed and the number of terminal nodes decreases by one. The same thing happens when there is three children nodes, except that the number of terminal nodes decreases by two. Because this move is the inverse of the growing step, the ratio within the minimum is simply the inverse of the ratio that is in the growing step.

3.7.3. Stay

When this step is chosen, it means that the present node become a terminal node. The probability is therefore:

$$\mathbb{P}_{\text{STAY}}(t) = r_t = 1 - b_t - d_t.$$

This is to respect the condition such as the sum of the three moves probabilities must be 1.

Denison and Mallick (2000) considers that the constant τ is as large as possible subject to

$$b_t + d_t \leq 0.75 \qquad t = 2, 3, ...$$

This is because we do not want too much moves in the algorithm. Furthermore, for $b = 1$, we put $b_t = 1$ because we want at least one split. Also, $d_t = 0$ if $b \leq 3$ for the same reason. We choose b_t so that it is twice as big as d_t when $\lambda = b + 1$. This is done in order to compensate for the fact that the birth step often proposes a tree which has fewer data points in the terminal nodes that is usually allowed. Furthermore, we want births and deaths to be proposed at a similar rate.

3.7.4. Specification of the algorithm

The algorithm used to develop the tree is given as follow. For a given node t that is on level L of the tree:

(1) examine every allowable split on each predictor variable and choose a splitting rule. The variable chosen is:

$$x_t = \underset{x_h \in (x_1, \ldots, x_p)}{\text{argmax}} \; q(x_h|y).$$

The split value chosen for this variable is:

$$s_l = \underset{s_m^{x_t} \in (x_t^{min}, x_t^{max-1})}{\text{argmax}} \; \pi_2(s_m^{x_t}|x_t, y).$$

(2) Set η equal to the number of terminal nodes in the present level of the tree ($\eta = 3^L$ as explained in section 3.2). This is because we want each

node on a same level to be split with the same probability.

(3) Generate $u \sim$ Multinomial such as:

$$u = \begin{cases} 1 & \text{with probability} \quad b_\eta, \\ 2 & \text{with probability} \quad r_\eta, \\ 3 & \text{with probability} \quad d_\eta. \end{cases}$$

(4) Use u to decide which of the three moves explained before is executed:

- if $u = 1$, grow the node, *i.e.* add two or three children nodes,
- if $u = 2$, prune the node, *i.e.* delete two or three children nodes,
- if $u = 3$, stay, *i.e.* make the node a terminal node.

(5) Repeat step 1 to 4 on the next node until there is no further split.

3.8. Assigning classes to terminal nodes

When the trees are constructed, each terminal node is assigned to one class. In Chapter 2, we saw that there is two classes in this project: the buyers and the non buyers, that is, each terminal node is considered either as a buyer, either as a non buyer. Hence each observation within the terminal node is assigned to its node category. In order to assign these classes, the target variable (closing ratio) is used. Lets define some terminologies:
Z:= the closing ratio on total population,
$z_j(i)$:= the closing ratio of the observation j in terminal node i,
$\bar{z}(i)$:= the average closing ratio in terminal node i,
$I(z_j|i)$:= the indicator of buying for observation j at node i.
The indicator $I(z_j|i)$ determines if an individual is considered as a buyer or not. If $I(z_j|i) = 1$, the observation j at node i is considered as a buyer.

Definition 3.8.1 (Class assignment rule in the context of K classes). *A class assignment rule assigns a class $k \in \{1, ..., K\}$ to every terminal node on the tree.*

Note that we are in a two-classes problem. Therefore, in each terminal node, we assign a class to every observation such as:

$$I(z_j|i) = \begin{cases} 0 & \text{if } z_j(i) < Z, \\ 1 & \text{otherwise.} \end{cases}$$

For $i = 1, ...b$, we determine if the observations in node i are considered as buyers with the function $H(z|i)$ as follow:

$$H(z|i) = \begin{cases} 0 & \text{if } \mathbb{P}(I(z|i) = 0) > \mathbb{P}(I(z|i) = 1), \\ 1 & \text{otherwise.} \end{cases}$$

where $\mathbb{P}(I(z|i) = 1) = \frac{1}{n_i} \sum_{j=1}^{n_i} I(z_j = 1|i)$ and $\mathbb{P}(I(z|i) = 0) = 1 - \mathbb{P}(I(z|i) = 1)$.

For example, in Figure 3.1 of Example 2.1.1, each node average closing ratio ($\bar{z}_i$) is written within the nodes. Therefore, because the population average closing ratio is $\bar{z}_N = 0.38$, it means that a terminal node will be considered as a buyer node if its average is greater than this value. Therefore, the only node with this characteristic is the node labeled 4 ($\bar{z}_4 = 0.88$). The observations in the other terminal nodes are thus considered as observations with a closing ratio of 0.

3.9. Criteria for tree selection

Many trees are constructed and then compared to identify trees of most interest. A "good tree" is a tree with the following characteristics:

(1) does not have too many terminal nodes,
(2) is formed with interesting variables in the business context,
(3) has terminal nodes with enough observations in it,
(4) has a large posterior probability,
(5) has a low misclassification rate.

Chipman and McCullogh (1998) presented many criteria to identify good trees.

3.9.1. Posterior probability of the tree

We define $\tilde{T}$ as the set of the trees constructed. If we form B different trees, this set is such as $\tilde{T} = \{T_1, ..., T_B\}$. For each tree T_h, we evaluate its posterior probability defined as

$$\mathbb{P}(T_h|\vec{y}) = \mathbb{P}(\vec{y}|\vec{x}, \vec{s}, T_h)\mathbb{P}(T_h) \qquad h = 1, ...B.$$

The tree with the largest posterior probability is chosen. However, Chipman and McCullogh (1998) discussed on a problem linked to the prior choice. Indeed, it is not the better approach when the prior on split value is too diffuse, like discussed in section 3.5.2. This problem is called the dilution effect. A criterion for tree selection that avoids this difficulty is to use the likelihood $\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T})$ as described below.

3.9.2. Likelihood $\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T})$

The likelihood $\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T})$ is independent of the dilution effect. Thus we can choose the tree T_h, such as :

$$T_h = \underset{T_h \in \tilde{T}}{\operatorname{argmax}} \, \mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T}).$$

An interesting criterion is also to plot the largest observed values of $\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T})$ against the number of terminal nodes of T. It allows us to quantify the value of adding terminal nodes while removing the influence of the tree prior (Chipman and McCullogh, 1998).

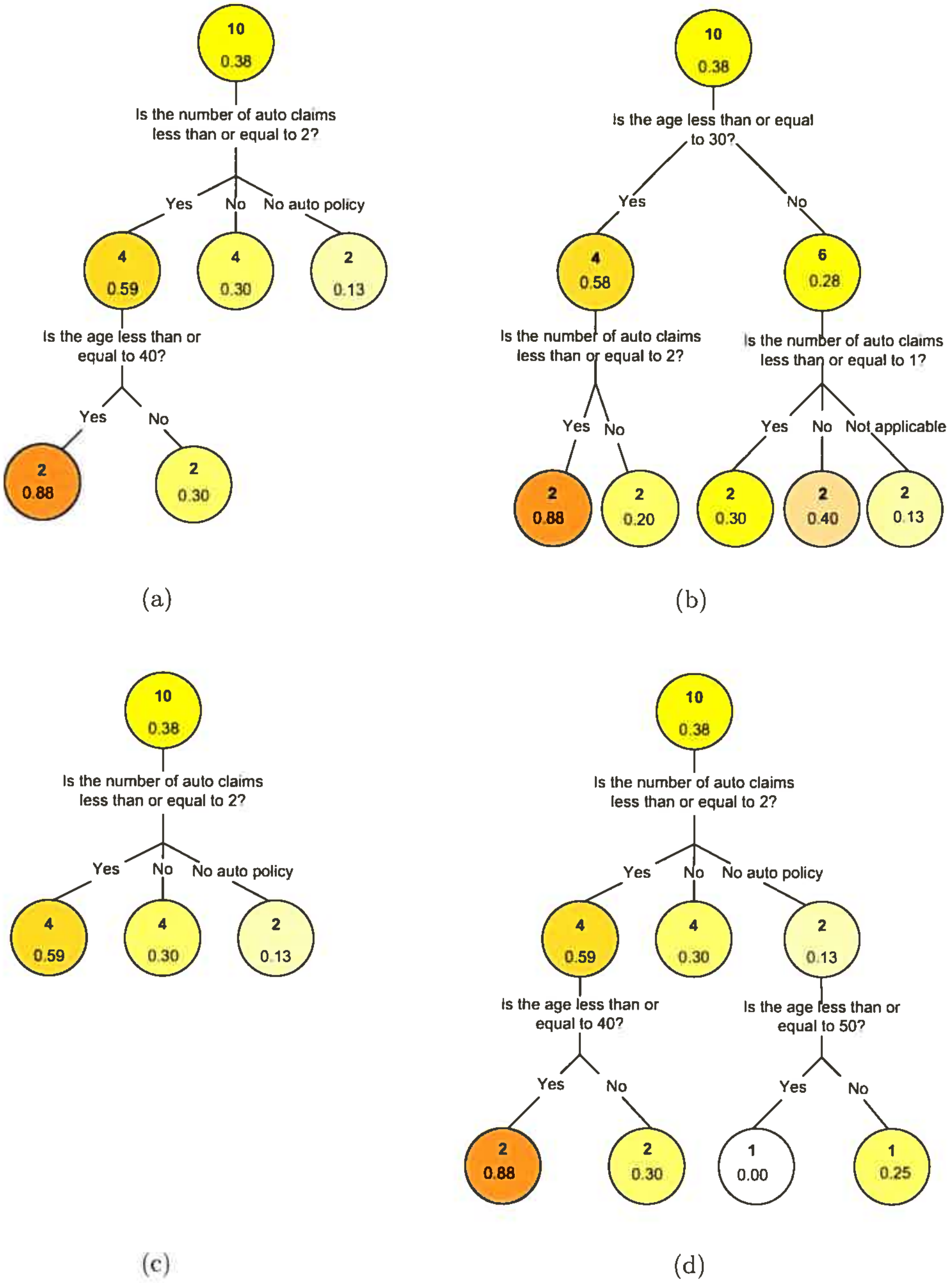
3.9.3. Misclassification rates

Another good criterion is the misclassification rate, denoted MIS_T. This refers to the total number of observations different from the majority at each terminal node. Therefore, we compare the two indicator functions presented in the previous section for all trees $T \in \tilde{T}$ as follow:

$$MIS_T = \frac{1}{n}\sum_{i=1}^{b}\sum_{j=1}^{n_i} |I(z_j|i) - H(z|i)|, \qquad (3.9.1)$$

where $n = \sum_{i=1}^{b} n_i$. With this criterion, we choose the tree such that:

$$T_h = \underset{T_h \in \tilde{T}}{\operatorname{argmin}} \, MIS_{T_h}.$$


10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Yes
No
2
0.88
2
0.30
10
0.38
Is the age less than or equal to 30?
Yes
No
4
0.58
6
0.28
Is the number of auto claims less than or equal to 2?
Yes
No
2
0.88
2
0.20
Is the number of auto claims less than or equal to 1?
Yes
No
Not applicable
2
0.30
2
0.40
2
0.13
10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
10
0.38
Is the number of auto claims less than or equal to 2?
Yes
No
No auto policy
4
0.59
4
0.30
2
0.13
Is the age less than or equal to 40?
Yes
No
2
0.88
2
0.30
Is the age less than or equal to 50?
Yes
No
1
0.00
1
0.25

(a)

(b)

(c)

(d)

FIG. 3.8. Representation of four trees using Example 2.1.1: (a) Tree T_1, (b) Tree T_2, (c) Tree T_3, (d) Tree T_4. The three criteria for the tree selection are calculated in order to find the best tree.

The tree assessment methods presented above are applied to four trees formed using data from Example 2.1.1. They are illustrated in Figure 3.8. The Table 3.1 shows that tree T_2 has the greater posterior probability and so the greater

TAB. 3.1. Tree assessments for Example 2.1.1 using a weight of 0.64 for the variable "age" and 0.36 for the variable "number of claims". The expected number of terminal node is 5. The best tree is indicated by the red color.

T_h	Posteriori	Likelihood	Misclassification rate
T_1	6.59×10^{-25}	2.31×10^{-10}	0.1
T_2	1.85×10^{-19}	2.53×10^{-07}	0.1
T_3	5.60×10^{-22}	7.60×10^{-10}	0.3
T_4	7.60×10^{-29}	4.26×10^{-13}	0.1

likelihood. Furthermore, its misclassification rate is also very low with a rate of 0.1. Although T_1 has also a misclassification rate of 0.1, its likelihood is smaller than T_2. T_3 has the worst misclassification rate among the four trees. The tree T_4 has the lowest likelihood even if its misclassification rate is low. Among the four trees, the tree T_2 is therefore chosen because it is the best for the three assessment methods.

In this chapter, we have presented a Bayesian method called BART to construct trees and calculate their posterior probability. The objective was to model the closing ratio as the target variable y. First, Section 3.1 presented the BART general model and some terminology on the tree. In order to calculate the tree posterior probability, two distributions were essential: the distribution of y and the tree prior. The distribution of y was seen as a binomial distribution depending on the purchases, the quotes and the predictor variables. Indeed, we wanted to calculate the closing ratio as the number of purchases among the quotes. To calculate the tree prior, three intermediate priors were used to obtain a tree with interesting business variables with not too many terminal nodes. Therefore, the variable prior was defined as a weight function of the variable importance. The prior for the splits given the chosen variable depended on the variable range. Finally, the prior for the number of terminal nodes was given as a truncated Poisson. Once these elements were known, the tree posterior probability was found.

This chapter explained also the three steps of the tree construction. On a certain node, the algorithm could split it, prune it or finally, decide to stop there. When the tree was constructed, we wanted to use it to predict the target variable, here the closing ratio. Therefore, we had to determine which terminal nodes were seen as buyers. Finally, the last step of the BART approach was to evaluate the trees and to find the most interesting one.

In the next chapter, the classical approach to construct the tree, called CART will be compared to the BART method by using the data described in Chapter 1.

Chapter 4

RESULTS

In the previous chapters, we present some techniques to better understand the Direct Market clients and prospects. Indeed the main objective of this project is to develop a model to predict the potential buyers of insurance products in the population under consideration. Therefore, the descriptive analysis and some insurance notions are exposed in Chapter 1. Chapter 2 includes two classical methods in statistic and it is divided in two distinct parts: one to better describe the population and the other to develop a predictive model of the closing ratio. Hence, the first section of Chapter 2 presents cluster analysis which allows to divide the population in homogeneous groups. The second section explains the "Classification And Regression Tree" (CART) method that forms decision trees used to predict if an individual will buy or not an insurance product. In Chapter 3, decision trees are also discussed but the approach used for the construction is very different from the CART method since it is a Bayesian approach (BART).

Therefore, this chapter presents the different results of these analyses. Even though the project is done for four regions in Canada, this chapter shows only the detailed results for the province of Ontario. Therefore, Section 4.1 covers the results for the classical methods which are the cluster analysis and the CART models. Section 4.2 describes the results under the Bayesian approach and the comparison between the CART and BART models is presented in Section 4.3.

4.1. Classical Results

4.1.1. Clustering

To divide the population into homogeneous groups, cluster analysis is applied to the data set. Because K-means clustering is an efficient method for large database, this method is chosen instead of hierarchical clustering. Furthermore, the statistical software SAS Enterprise Miner 5.2 allows to compute this clustering technique. It calculates the optimal number of clusters using the CCC criterion explained in Section 2.1.6. As the descriptive analysis, the results presented here are for the province of Ontario.

In Ontario, among the 105,028 individuals, 7 clusters are found using the centroid distance. The distribution of these clusters is represented in Figure 4.1.

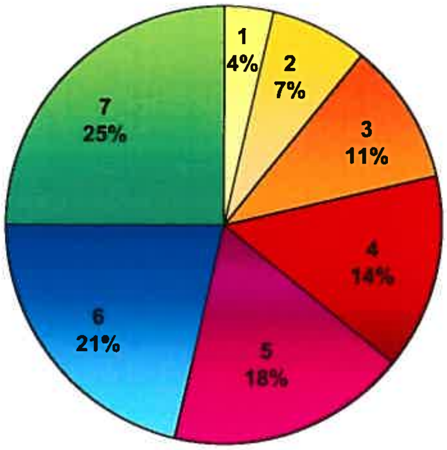


FIG. 4.1. Ontario cluster distribution in 2005 using the K-means algorithm with the centroid distance.

4.1.1.1. *Description*

The clusters are described in Tables 4.1 and 4.2 using many characteristics. Each column refers to one of the 7 clusters. They are presented in order of percentage of the population they contained. For the automobile variables, the percentages are not for the entire population but for the individuals who have or asked for an auto policy. For example, in cluster 2, among the 86% of persons who have or asked for an auto policy, 67% have their license since more than 10 years. The same principle is applied to the residential variables. Thus, in cluster 2, among the 45% of the persons who have or asked for a residential policy, 71% have or asked for a homeowner package. Furthermore, we note that in clusters 1,4 and 5, the percentages of individuals who have or asked for residential policy is almost null.

4.1.2. Decision tree

An important consideration in the closing ratio model is the impact of the clusters formed previously. Indeed, at the beginning of the project, the purpose was to use the cluster variable in the tree to see which clusters are considered as buyers and which are not. However, it was decided not to use it because the clusters do not include the same proportion of clients and prospects. Consequently, the clusters with a large proportion of clients will have a good closing ratio and those with a large percentage of prospects will have a very low closing. Indeed the prospects are individuals that did not buy the insurance products.

To obtain the best closing ratio model as possible, many trees are developed using SAS Enterprise Miner 5.2. This software allows to enter the following parameters:

(1) the impurity function,
(2) the minimum number of observations included in a node.
(3) the maximum number of tree level.

TAB. 4.1. Percentage of observations with each characteristic in the Ontario main three clusters. Only significant variables are presented.

Variables	Cluster 2	Cluster 1	Cluster 5
Percentage of the population	27	25	18
Account since more than 10 years	96	27	53
Age less than 30	9	27	26
Age between 45 and 65	36	26	23
Have or asked for an Auto policy	86	99	99
Have or asked for collision protection	85	98	2
Have or asked for condo package	14	<1	<1
Have or asked for homeowner package	71	<1	<1
Have or asked for a residential policy	45	2	0
Have or asked for tenant package	14	<1	<1
License since more than 10 years	67	44	24
Male	66	68	83
Private passenger vehicle	100	94	94
Prospects	16	88	43
Vehicle age less than 4 years	32	51	1
Closing ratio in percentage per cluster	10	1	34

4.1.2.1. *Impurity*

Many trees are constructed using the different impurity functions exposed in Section 2.2.5. Then, the tree with the smaller misclassification rate is chosen. In practice, the best results were found with the entropy function.

4.1.2.2. *Stopping rule*

To determine the splitting rule, both methods of Section 2.2.7 are used. First, when a split stops producing child nodes, it is because the closing ratio is not different enough. This method corresponds to the minimal change in impurity. Furthermore, the minimal number of observations is included in the parameters.

TAB. 4.2. Percentage of observations with each characteristic in the remaining Ontario clusters. Only significant variables are presented.

Variables	Cluster 6	Cluster 4	Cluster 7	Cluster 3
Percentage of the population	13	8	6	3
Account since more than 10 years	46	62	29	33
Age less than 30	21	10	19	24
Age between 45 and 65	25	40	29	31
Have or asked for an Auto policy	91	98	12	91
Have or asked for collision protection	67	33	58	89
Have or asked for condo package	15	<1	8	1
Have or asked for homeowner package	43	<1	33	97
Have or asked for a residential policy	45	3	87	43
Have or asked for tenant package	28	<1	25	3
License since more than 10 years	32	80	70	43
Male	67	87	64	76
Private passenger vehicle	94	11	0	100
Prospects	15	17	86	92
Vehicle age less than 4 years	27	42	41	41
Closing ratio in percentage per cluster	80	60	1	3

Consequently, when a node is too small, it is not divided and is therefore a terminal node. The threshold for the minimum number of individuals in a node is usually fixed at 10 % of the total number of clients and prospects in the data set.

4.1.2.3. *Splitting rule*

Under the classical approach, two techniques are used in the choice of the splitting rule. In the first method, a tree is constructed using an automatic way. Indeed, the starting parameters presented above are first determined and each node is splitted using the splitting rule that minimize the impurity function.

This algorithm continues until a stopping rule is reached. Under this technique, each variable and split is chosen without regards of its interpretability. However, choosing always the splitting rule that minimizes the impurity does not mean that the final tree is optimal. Indeed the optimal tree is the tree with the overall lowest impurity and it can be obtained in many ways.

An alternative technique consists to add some *a priori* information to the classical method. It is thus a combination between a classical and a Bayesian method. It will be referred to as the combination method. When the node is splitted under this approach, it is not always the optimal splitting rule that is chosen. Indeed, the change in impurity is calculated for each possible splitting rule and the statistician chooses the most interesting one for the business.

For each technique, many trees are constructed using different parameters. The next subsections present the chosen tree for each method.

4.1.2.4. *Tree from the pure classical approach*

The final tree using the pure classical approach for the splitting rule is presented in Figures 4.2 to 4.4. The root node is divided into three child nodes using the variable "collision coverage". The left and middle child nodes are subtrees while the right child is not splitted. From this tree, it can be seen that the individuals without the collision coverage are those with the largest closing ratio. Figure 4.2 also shows that the persons that do not have automobile information do not have a good closing ratio. Indeed the closing ratio in the node is below the population average (0.28).

Figure 4.3 shows that among the individuals who have the collision coverage, those with a closing ratio of 0.60 (see the number 1 in this figure) have the following characteristics:

- no residential information and,
- at least one snowmobile vehicle and,

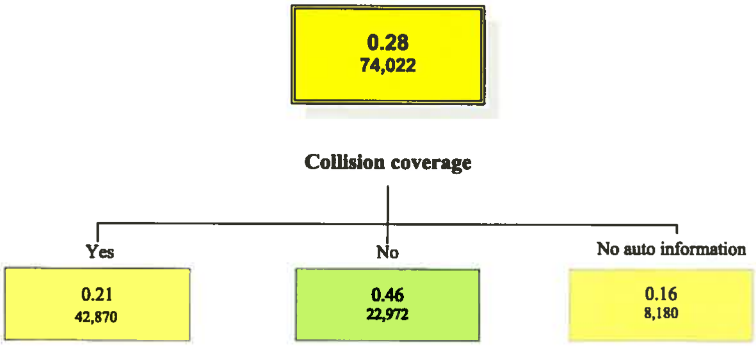


FIG. 4.2. Root node and first level of the CART model for Ontario using the pure classical approach and the entropy impurity function.

- an account since more than 8.5 years.

We also see that the persons without an homeowner package have a closing ratio of 0.56 (see the number 2 in this figure). Furthermore, the subtree with the individuals that have a collision coverage has 6 levels and 8 terminal nodes.

Figure 4.4 presents the middle subtree of Figure 4.2. It therefore includes individuals without the collision coverage and among them, two groups have a closing ratio that is more than 0.50 (see the two numbered nodes in this figure). Their characteristics are the following:

- a license since more than 3 years, an account since more than 8.5 years and at least one snowmobile vehicle (see the number 1 in this figure),
- a license since less than 3 years and no motorcycle vehicle (see the number 2 in this figure).

In conclusion, the whole tree under the classical approach is presented in Figure 4.5. This tree has therefore a total of 14 terminal nodes and 7 levels.

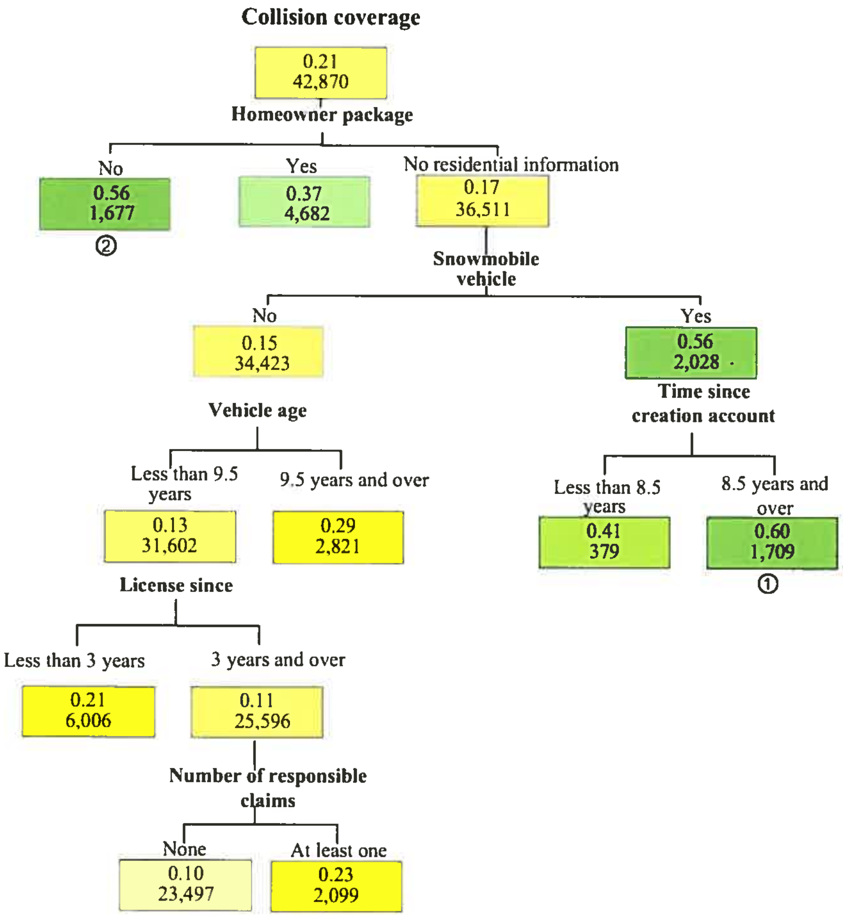


FIG. 4.3. Subtree of the Ontario CART model with individuals that have a collision coverage. The pure classical approach is used with an the entropy impurity function. The two numbers identify the nodes with the biggest closing ratios. This subtree has 6 levels and 8 terminal nodes.

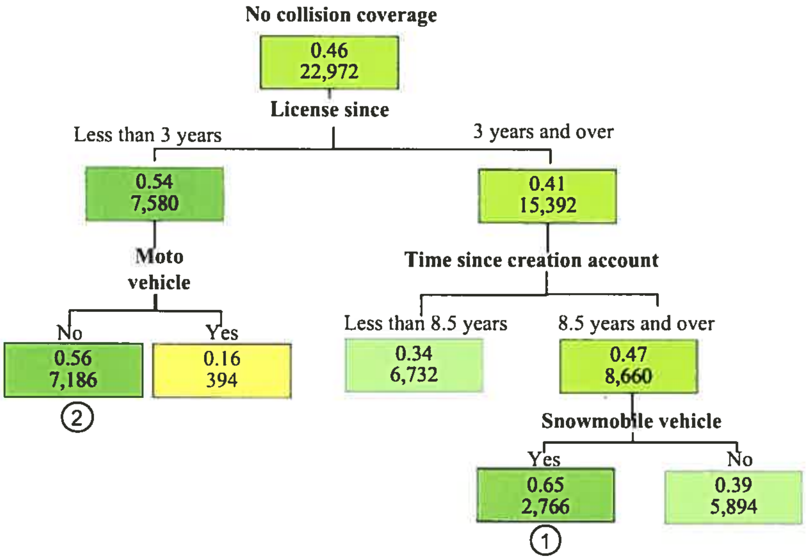


FIG. 4.4. Subtree of the Ontario CART model with individuals that do not have a collision coverage. The pure classical approach is used with an the entropy impurity function. The two numbered nodes are those with the biggest closing ratios. This subtree has 4 levels and 5 terminal nodes.

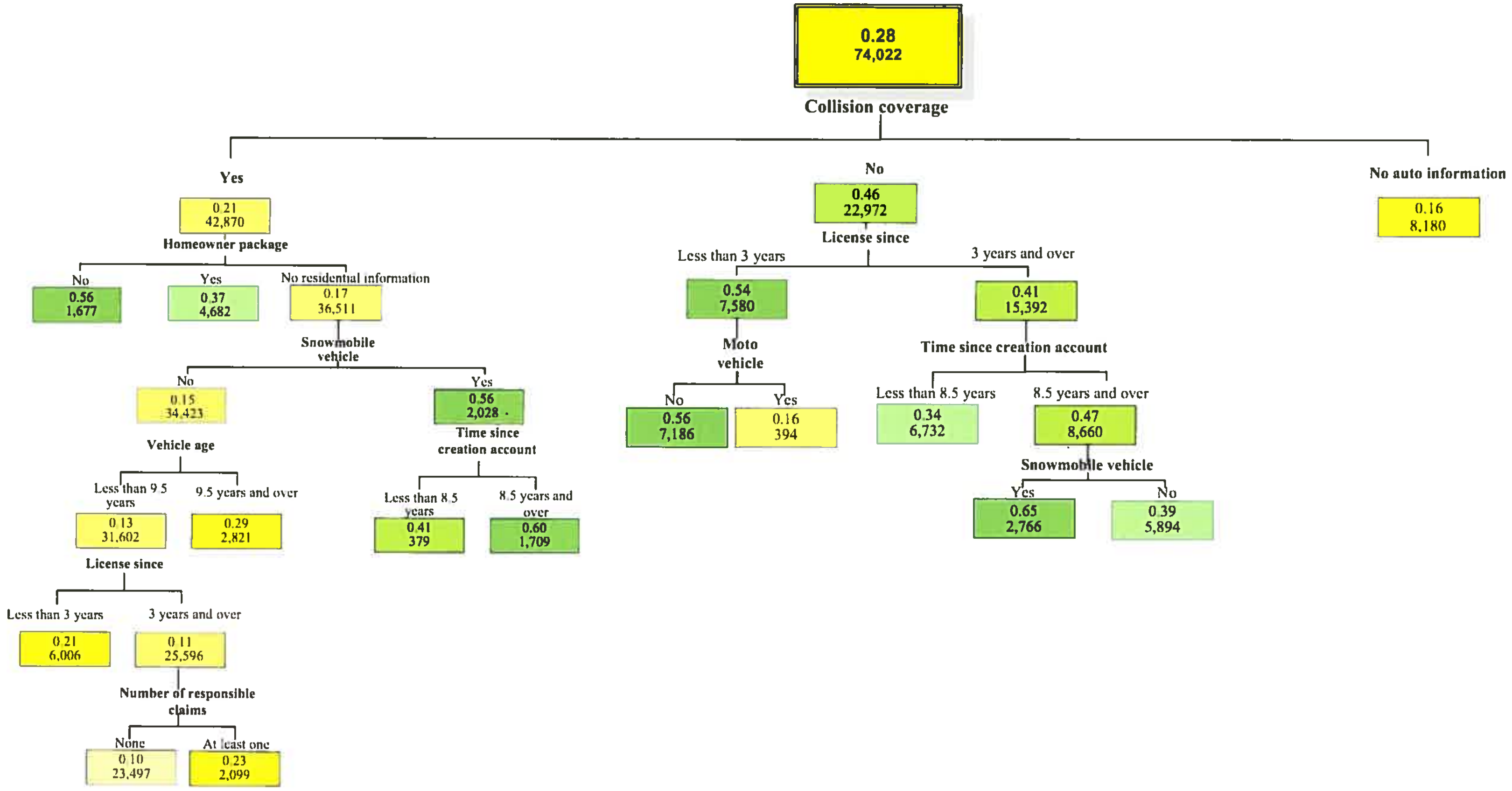


FIG. 4.5. CART model for Ontario using the pure classical approach and the entropy impurity function. There is a total of 14 terminal nodes and 7 levels.

4.1.2.5. *Tree from the combination approach*

In the method that combines the classical and Bayesian approaches, the first split variable is chosen for business consideration. Indeed, for the clients (the "Business strategies" department), the first variable in the model is very important and it is better if it is a demographic variable. Therefore, the first split variable is the "age" for all regions. These separation allows to focus on distinct group ages.

Figure 4.6 shows the global tree for Ontario that is produced with this approach using the entropy as the impurity function. In this model, the first split divides the root node in three subtrees: one for the individuals younger than 30 years (see Figure 4.7), one for those between 30 and 55 years old (see Figure 4.9) and the last one with the persons older than 55 years old (see Figure 4.8). The whole tree is presented in Figure 4.10. In the entire population, the closing ratio is 0.28 and the tree allows to find groups with an average closing twice this value.

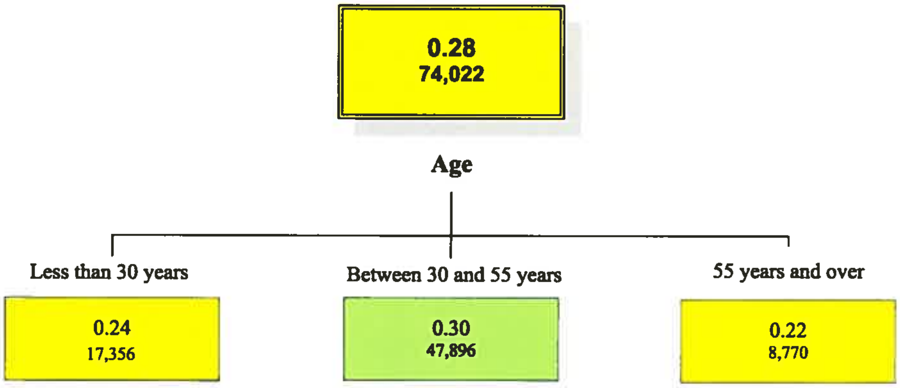


FIG. 4.6. Root node and first level of the CART model for Ontario using the combination between the classical approach and some *a priori* information. The impurity function is the entropy function.

Figure 4.7 presents that among the persons younger than 30 years old, the individuals with the biggest closing ratio (see number 1 in the figure) have the collision coverage, a vehicle age less than 8 years and a snowmobile vehicle.

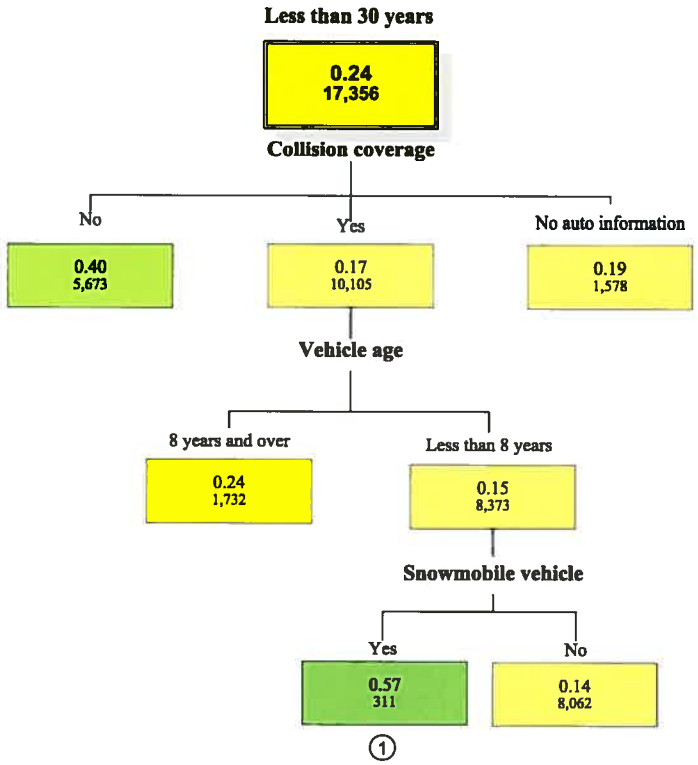


FIG. 4.7. CART model for the individuals younger than 30 years old in Ontario using the combination between the classical approach and the addition of *a priori* information. The impurity function is the entropy function and the subtree has 4 levels and 5 terminal nodes.

Figure 4.8 presents the clients and prospects older than 55 years old. Among them, two groups have a closing ratio of 0.55:

(1) the persons without a collision coverage and without a private passenger vehicle (see number 1 in this figure),

(2) the persons with a collision coverage, a license since more than 20 years and a snowmobile vehicle (see number 2 in this figure).

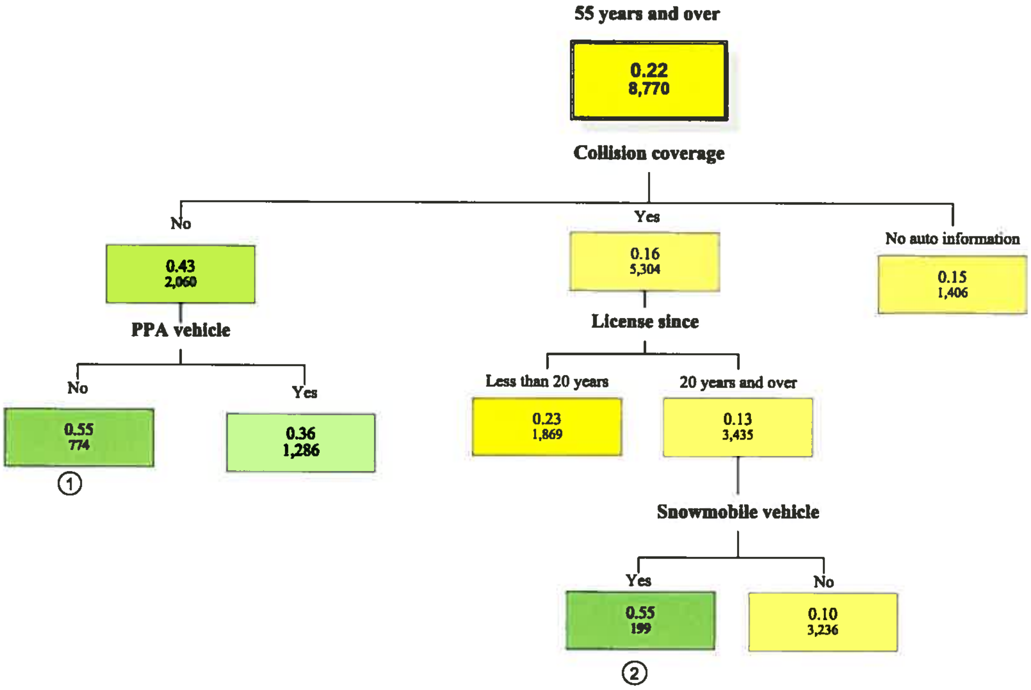


FIG. 4.8. CART model for the individuals older than 55 years old in Ontario using the combination between the classical approach and some *a priori* information. The impurity function is the entropy function. The two numbered nodes are those with the biggest closing ratios (0.55). Furthermore, this subtree has 6 terminal nodes and 4 levels.

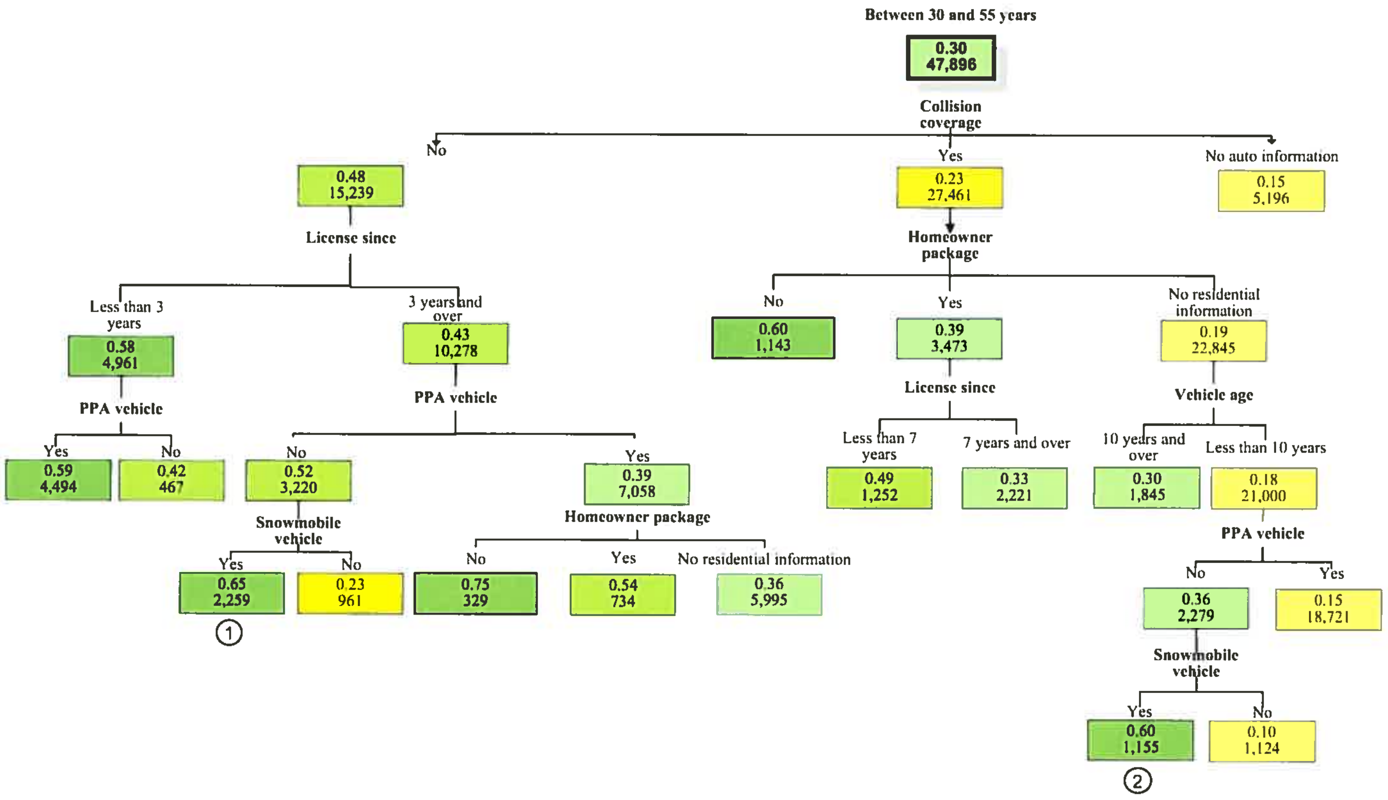


FIG. 4.9. CART model for the individuals between 30 and 55 years old in Ontario using the combination between the classical approach and some *a priori* information. The impurity function is the entropy function. Furthermore, this subtree has 14 terminal nodes and 6 levels.

Figure 4.9 illustrates the subtree of the persons between 30 and 55 years old. Among them, we see that the individuals with a closing ratio of 0.65 (see the number 1 in the figure) do not have the collision coverage, a license since more than 3 years, no private passenger vehicle and at least one snowmobile vehicle. Furthermore, the individuals that have a collision coverage, no residential information, a vehicle age less than 10 years, no private passenger vehicle and a snowmobile vehicle have a closing ratio of 0.60 (see the number 2 in Figure 4.9).

The whole tree under the combination approach is illustrated in Figure 4.10 and it allows to see that the whole tree has a total of 25 terminal nodes and 7 levels.

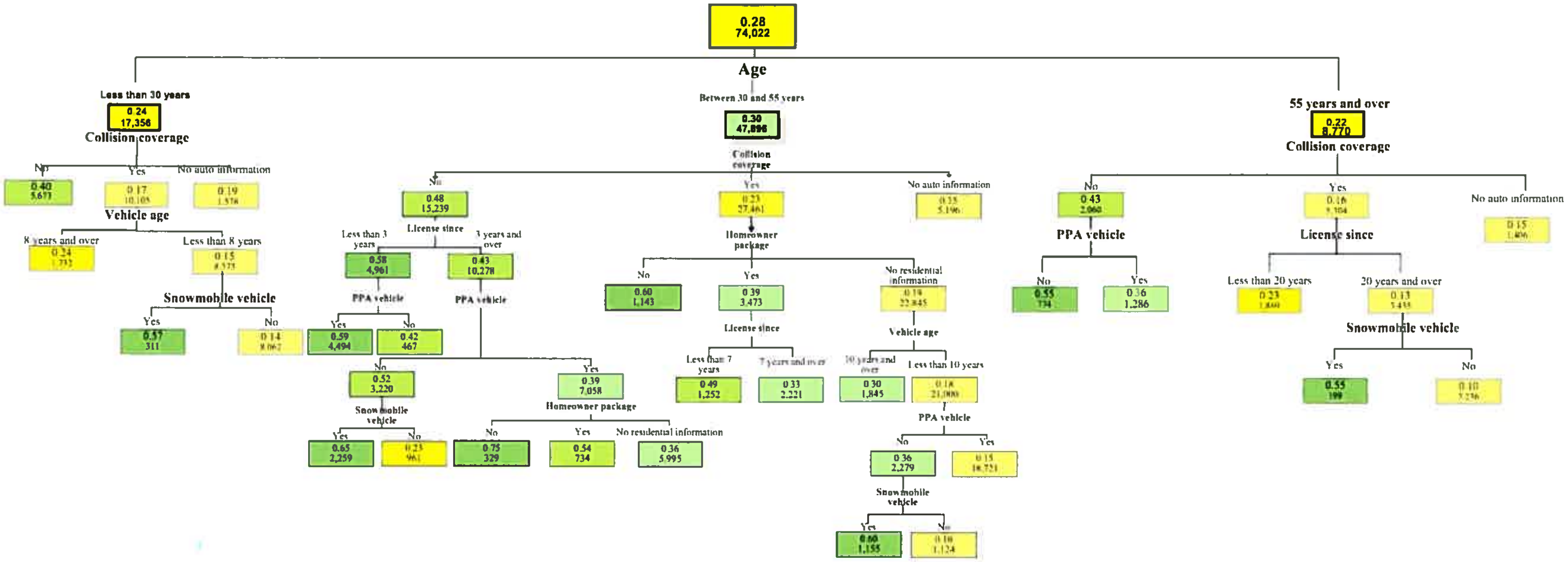


FIG. 4.10. CART model for Ontario using the combination between the classical approach and some *a priori* information. The impurity function is the entropy function and the tree includes 25 terminal nodes and 7 levels.

4.1.2.6. *Comparison between the two classical techniques*

Now that the two constructed trees are presented, we want to know what approach is the best between them. The two techniques show similar trees in term of closing ratio. Indeed, both methods allow to find groups with a closing ratio that is twice as the population average. Although the significative variables are very similar, the tree shapes are different. Indeed, Table 4.3 shows the number of terminal nodes and the total number of levels for the two approaches. It is possible to see that the combination method produces more terminal nodes with one less level.

TAB. 4.3. Description of the tree shapes using the two different classical approach for the construction of CART models.

Shape	Pure classical	Combination method
Number of terminal nodes	13	25
Number of levels	7	6

To determine the method that gives the best result, the misclassification rates are calculated and presented in Section 4.3.

4.2. BAYESIAN RESULTS

Decision trees are also constructed using the Bayesian approach (BART) that is, the theory that is covered in Chapter 3 is applied and the results are presented in this section. Because the first step is the choice of the splitting rule, the following items must be determined for each variable before the beginning of the algorithm:

(1) the score which also determines each variable weight calculated with equation (3.5.2) of Section 3.5.1,
(2) the variable distribution (see Section 3.3),
(3) the set of possible splits (see Subsection 3.5.2).

Table 4.4 presents each scores and weights for the predictor variables. The variable with the highest importance is "collision protection". Table 4.5 shows the predictor variables distributions (see Section 3.3).

TAB. 4.4. Predictor variable scores and weights.

Predictor variables	Score	Weight
Age	7	0,0522
Collision deductible	10	0,0746
Gender	4	0,0299
High performance vehicle	1	0,0075
Household average income	3	0,0224
Indicator of automobile information	8	0,0597
Indicator of residential information	9	0,0672
Licence since	8	0,0597
Newest vehicle age	8	0,0597
Number of All-Terrain-Vehicles	6	0,0448
Number of comprehensive claims	1	0,0075
Number of condo packages	5	0,0373
Number of homeowner packages	9	0,0672
Number of moto vehicles	3	0,0224
Number of non responsible collision claims	4	0,0299
Number of other vehicles	2	0,0149
Number of private passenger vehicles	7	0,0522
Number of responsible collision claims	4	0,0299
Number of snowmobiles	7	0,0522
Number of tenant packages	5	0,0373
Number of years since the last accident	7	0,0522
Oldest vehicle age	8	0,0597
Renting vehicle	1	0,0075
Time since creation account	6	0,0448
Vehicle credit	1	0,0075

TAB. 4.5. Predictor variable distributions.

Predictor variables	Distribution
Age	gamma
Collision deductible	binomial
Gender	binomial
High performance vehicle	binomial
Household average income	normal
Indicator of automobile information	binomial
Indicator of residential information	binomial
Licence since	gamma
Newest vehicle age	gamma
Number of All-Terrain-Vehicles vehicle	Poisson
Number of comprehensive claims	Poisson
Number of condo packages	Poisson
Number of homeowner packages	Poisson
Number of moto vehicles	Poisson
Number of non responsible collision claims	Poisson
Number of other vehicles	Poisson
Number of private passenger vehicles	Poisson
Number of responsible collision claims	Poisson
Number of snowmobiles	Poisson
Number of tenant packages	Poisson
Number of years since the last accident	Poisson
Oldest vehicle age	gamma
Renting vehicle	binomial
Time since creation account	mixture of gamma
Vehicle credit	binomial

Before the beginning of the BART construction algorithm, some parameters must be determined.

- **Average number of terminal nodes**: this is the desired number of terminal nodes in the tree and it is given by $b = E(\eta)$. Therefore, it will determine the λ parameter that is used for the birth and pruning probabilities. The average number of terminal nodes wanted depends on the number of observations in the total population. Because the terminal nodes must include enough observations, a region that contains a big number of individuals will have more terminal nodes.

- **Minimum number of observations**: this parameter is the minimum number of individuals that a node can have in order to be splitted. When a node contains a small number of observations, the algorithm forces it to stay as terminal node. We fixe the threshold around 10% of the total number of observations.

- **Parameter** τ: this parameter is used in the birth and pruning probabilities. In line with the constraint $b_t + d_t \leq 0.75$ (see Section 3.7), many possible values are tested between 0.15 and 0.2.

These parameters are different depending on the region model. Hence, for each region, many different values were tried for each parameter and the best combination was kept. Consequently, many trees are formed for each region to obtain the best possible model. Then, the two following criteria (see Chapter 3) are calculated:

- the log-likelihood $\log(\mathbb{P}(\vec{y}|\vec{x}, \vec{s}, \vec{T}))$,
- the misclassification rate.

To obtain the likelihood, all the observation densities are multiplied together. However, when we multiply n small numbers, the result is near 0. This is why the logarithm of this function is used. Furthermore, even with this transformation, the constant $c = \log(10^{300})$ is added to the log-likelihood because of the large number of observations in the data sets (74,022 in Ontario).

These two criteria are thus computed for many models. For each region, the results of two combinations of parameters are presented. Hence, Table 4.6 shows the parameters that produce the best models for each region. On the other hand, Table 4.7 shows the parameters that are used for the BART models not chosen.

TAB. 4.6. Chosen parameters for the BART algorithm 1.

Region	Average number of terminal nodes	Minimum number of observations	τ
Ontario	20	6,000	0.15
Quebec	20	3,000	0.15
Western	15	3,000	0.20
Atlantic	10	200	0.20

TAB. 4.7. Chosen parameters for the BART algorithm 2.

Region	Average number of terminal nodes	Minimum number of observations	τ
Ontario	34	1,000	0.20
Quebec	24	4,000	0.20
Western	16	800	0.20
Atlantic	12	300	0.15

Furthermore, Table 4.8 presents the calculated criteria for each model. It is then possible to see that the first model has the smallest misclassification rate and the largest likelihood for all regions. Hence, the chosen models are those with the characteristics presented in Table 4.6.

For the province of Ontario, the final tree is presented with the Figures 4.11, 4.12 and 4.13. The complete tree is in Figure 4.14. Therefore, the first chosen variable is "snowmobile vehicle". It allows to find at the first level a group with an average closing ratio of 0.61 (see Figure 4.11). Indeed the individuals with a

TAB. 4.8. Assessment of the two Bart models for the four regions.

Region	Model 1		Model 2	
	log-likelihood	Misclassification rate	log-likelihood	Misclassification rate
Ontario	32,968	0.2653	2,125	0.2741
Quebec	14,378	0.1601	523	0.1614
Western	16,125	0.2655	14,511	0.2684
Atlantic	9,548	0.2672	6,239	0.2946

snowmobile vehicle have a large closing ratio. The three child nodes produced by this first split are defined as:

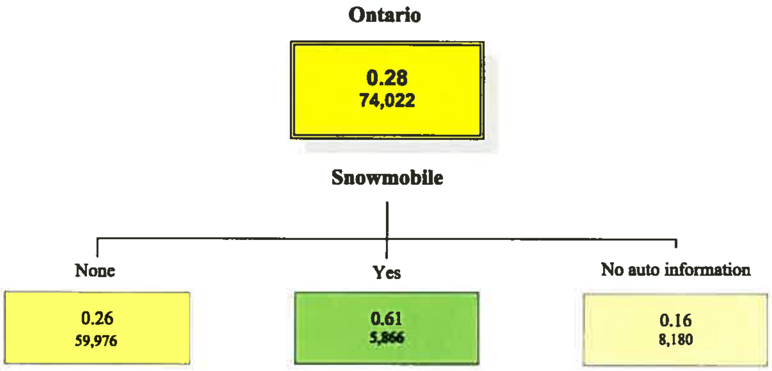


FIG. 4.11. Root node and first level of the BART model for the individuals in Ontario with τ=0.15 and an average number of terminal node of 20. The first variable that divides the total data set is the variable "snowmobile vehicle" which produces threes subtrees.

- The **left subtree** includes individuals without a snowmobile vehicle (see Figure 4.13). Among these persons, the number of private passenger vehicles, the licence since and the presence of residential information are important factors in the prediction of the closing ratio. This subtree has 12 terminal nodes on 6 levels.
- The **middle node** includes individuals with a snowmobile vehicle and is not splitted (see Figure 4.11). Indeed this node has a strong average

closing ratio and do not have enough observations to be divided. It is thus a terminal node.

- The **right subtree** contains persons without auto information and is only splitted one time. Consequently, there is two terminal nodes that have small closing ratios. (see Figure 4.12).

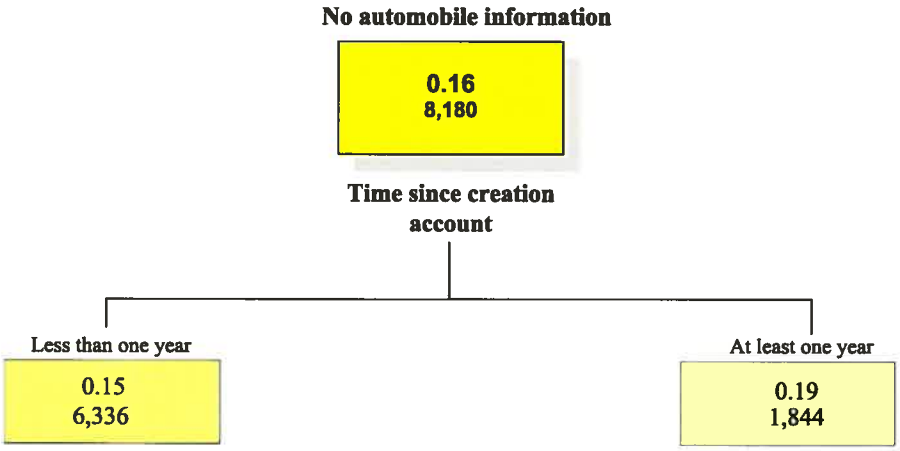


FIG. 4.12. BART model for the individuals of Ontario without automobile information available (right subtree). The model uses τ=0.15 and an average number of terminal node of 20. This subtree shows 2 terminal nodes and 6 levels.

The whole BART model is presented in Figure 4.14 and it includes 15 terminal nodes with a total of 7 levels. Furthermore, the most significant variables under this model were the number of "Private Passenger vehicles", the licence since and the presence of residential information.

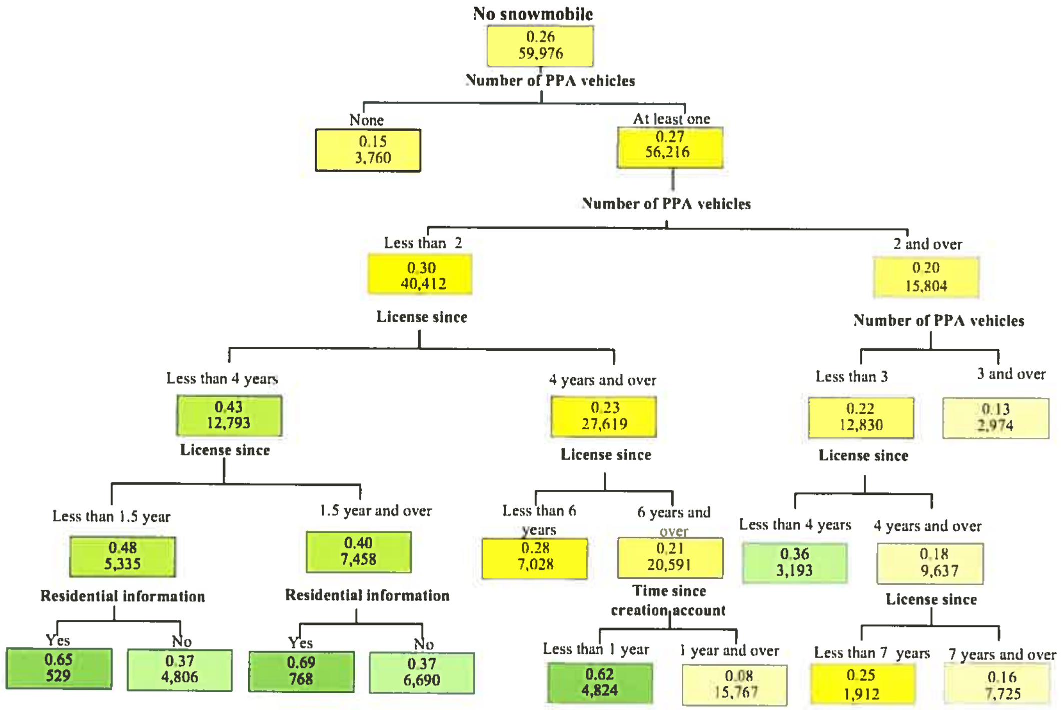


FIG. 4.13. BART model for the individuals with no snowmobile in Ontario (left subtree) with τ=0.15 and the average number of terminal node of 20. This subtree shows 12 terminal nodes and 6 levels.

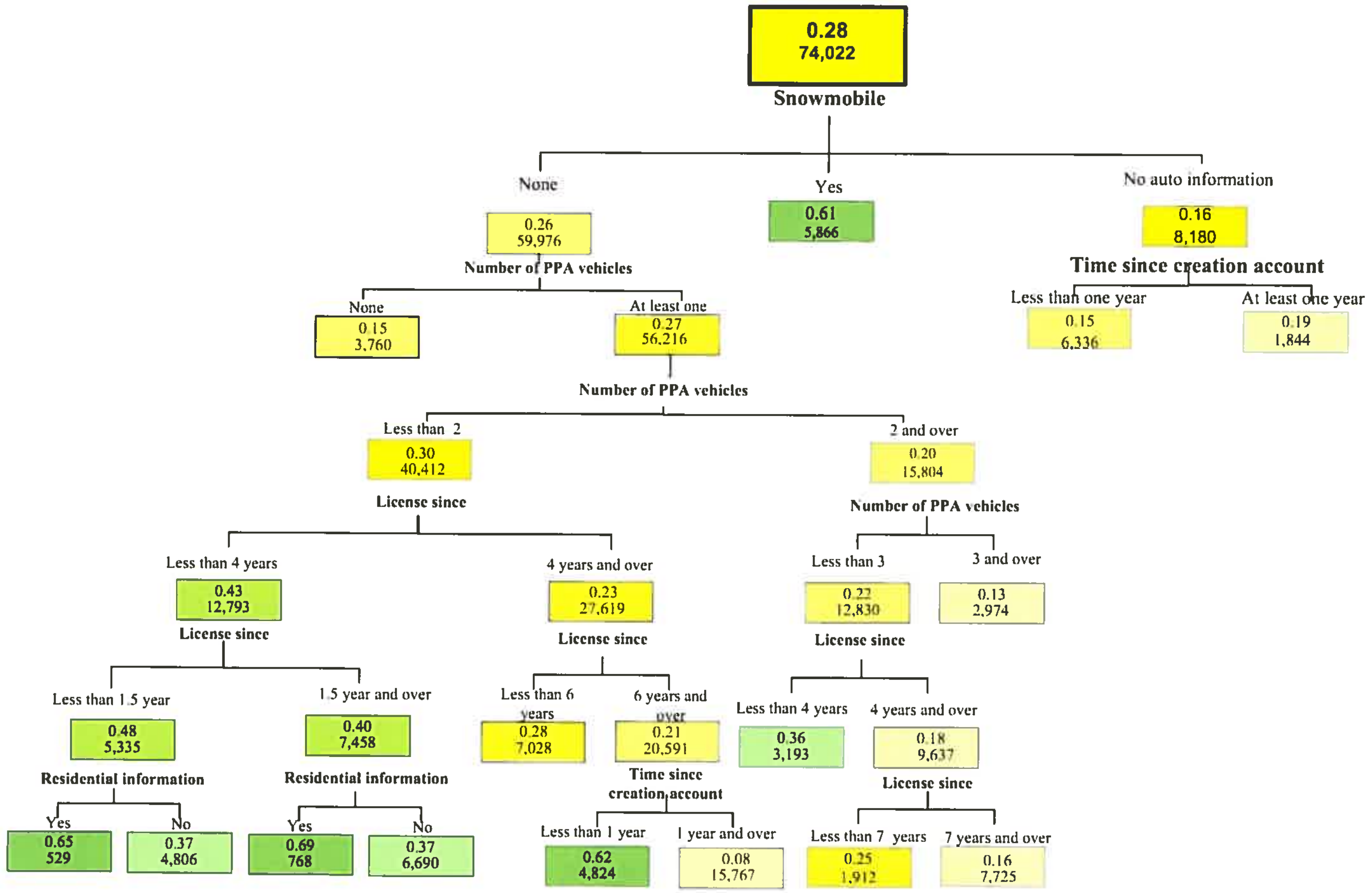


FIG. 4.14. BART model for the individuals in Ontario with $\tau=0.15$ and an average number of terminal node of 20. The model includes 15 terminal nodes with a total of 7 levels.

4.3. Comparison between classical and Bayesian approaches

Now that CART and BART models have been developed, it is possible to compare these approaches. In term of average closing ratio, the classical, the Bayesian and the combination of both approaches find nodes with average closing ratio that are as twice as the population average. Furthermore, the three trees have a total of seven levels.

The most significant variables in the three models are:

- the presence of snowmobile vehicle,
- the presence of residential information,
- the licence since,
- the time since the creation account.

However, the "collision coverage" variable is very important in the classical model but is not significant in the Bayesian model.

4.3.1. Interpretability

To evaluate each tree interpretability, the significant variables they contains are examined. Indeed, an interpretable tree includes important variables in a business matter. The score variables exposed in Table 4.4 are therefore used to calculate each tree average score. The nearest to 10 is the average score, the more interpretable the tree is. To calculate this measure, all variable splits are considered with their corresponding scores. For example, if the variable split is "number of snowmobiles", the corresponding score is 7. In a given tree, the average score is:

$$\bar{a} = \frac{1}{N_S} \sum_{s=1}^{N_S} a_s,$$

where a_s is the score of the variable x_s used to split and N_S is the total number of splits in the tree. For instance, there is thirteen splits in the BART model. Table 4.9 presents the average scores for the three models. It is possible to see that the CART combination includes more important variables because it has the best average score. Furthermore, the BART model has a similar average score while the pure CART model has the lowest score. However the BART model is much less complex than the combination model with a total of thirteen splits instead of nineteen. Table 4.9 also indicates that the pure CART model gives a tree with only eleven splits. Therefore, this method shows interesting results in term of interpretability because it gives a simple tree. Thus, depending of the tradeoff (score or number of splits), any of the three candidates could be chosen. However, because of his simplicity and his good average score, the BART model is seen as a good compromise between the two other approaches.

TAB. 4.9. Average score and number of splits for the three methods for the Ontario models.

Method	Average score	Number of splits
CART pure	6.9	11
CART combination	7.9	19
BART combination	7.6	13

4.3.2. Misclassification rates

To evaluate the tree performances, an other method is to calculate the misclassification rates for each model. Table 4.10 presents the results for the three different approaches. Therefore, in term of misclassification rate, the BART model is equivalent to the combination method. However, the result for the CART pure approach is not very different.

TAB. 4.10. Misclassification rates for decision trees. The red color indicates the best model.

Region	CART pure	CART combination	BART
Ontario	0.2674	0.2653	0.2653

In conclusion, for the province of Ontario, the Bayesian approach can be seen as a good compromise between the three approaches. Indeed, interesting results are found in term of interpretability. For the misclassification rate criterion, it shows the same result as the combination method. However, in opposition to the combination approach, the Bayesian approach allows to automatically construct the trees.

In this chapter, the detailed results were presented for the region of Ontario. In the cluster analysis, seven clusters were found by using the K-means algorithm with the centroid distance. For the closing ratio models, two classical approaches were presented: one that is a pure classical method and another that is a combination between the classical and the Bayesian approach. The predictive model was also developed under the Bayesian approach and we saw that the combination method and the Bayesian approach gave better results in term of misclassification rate. However, because the Bayesian model gave a tree simple and interpretable, it was considered as the method with the best overall performance.

CONCLUSION

In the context of insurance, this Masters thesis explores classical and Bayesian approaches to model nonlinear data. The project provides answers to two specific questions asked by TD Meloche Monnex.

(1) What kind of individuals are in the Direct Market population?

(2) In this population, who are the buyers and how we can predict them?

To answer the first question, the project starts with a descriptive analysis. This allows to see that the majority of the population under consideration is in the province of Ontario. Therefore, all detailed results are shown for this province. After the general data exploration, the cluster analysis is done to form different groups that are more homogeneous than the total population. Many clustering algorithms are presented but the best method in our context is the partitional clustering using the K-means algorithm. This is therefore computed using the statistical software SAS Enterprise Miner. In Ontario, seven clusters are thus formed, each with specific characteristics. This allows to identify which groups could be targeted in term of marketing.

To find the buyer characteristics, a statistical model that uses the closing ratio as the target variable is created. To construct this model, many different methods could have been used. Indeed, the target variable is binary and consists of determining if an individual is a buyer or not. Consequently, a logistic regression could be an interesting approach. However, the clients who asked for this project are not statisticians. Furthermore, the model interpretability is almost as important as its performance. In this context, the statistical technique that is chosen is the decision tree models.

To construct the decision trees, the classical and the Bayesian approaches are used. The classical approach, that is most often used in industry, is computed using the software SAS Enterprise Miner. Furthermore, two different methods are compared under this approach. The first method that is purely classical consists of always choosing the best splitting rule without regard to the *a priori* information. The formed trees are showing good performance in term of closing ratio because it allows to find groups with a closing as twice as the closing in the Direct Market population. However, this approach can be improved by adding *a priori* information. Indeed, a second method consists of using the same construction algorithm but with the choice of a splitting rule that takes into account the *a priori* information on variables. This approach shows models with smaller misclassification rates and also a good performance in term of closing ratio.

The other method to create the trees is the Bayesian approach. In this case, the models are called BART models. This technique allows to put more weight on variables that are important for business. Therefore, an algorithm exposed by Chipman and McCullogh (1998) is applied. With the objective of finding a tree with the maximum *a posteriori* probability, many trees are formed with different parameters. To compete many models, the likelihood and the misclassification rate are calculated on each possible tree. Indeed the best tree must have the smallest misclassification rate and the largest likelihood. Therefore the results show that the Bayesian approach give interesting models in terms of variables, misclassification rate and closing ratio.

In the Bayesian approach, the tree forest could have been used to construct a large number of trees. Indeed the tree forest is a sequence of trees that allows to combine many trees in the same model. Although this method could be more difficult to explain, it could help to reduce the misclassification rate.

In conclusion, the three methods used to construct the CART and BART models show interesting trees in term of closing ratio and tree shapes. Furthermore, the most significant variables are similar in the three models. Indeed, the most potential buyers have one or both of the following characteristics: snowmobile vehicle and residential quote or policy. Under the classical approach, the presence of collision coverage is an indicator of a smaller closing ratio.

To determine which approach is best, the tree interpretability and misclassification rate are examined for each model. Therefore, the BART model is more interpretable and is either better or equivalent to the combination approach in term of misclassification rate. The Bayesian approach is thus the method that gives the best results in this project.

Bibliography

ANDERSBERG, M. R. (1973). *Cluster analysis for applications.* Academic Press, New York.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and regression trees.* Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, CA.

BROWN, G. L. R., ROBERT L. (2001). *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance.* ACTEX Publications, Winsted, Connecticut.

CALIŃSKI, T. and HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communication in Statistics*, **3**, 1–27.

CHIPMAN, G. E., H. and MCCULLOCH, R. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, **93**, 935–960.

CHIPMAN, G. E., H. and MCCULLOCH, R. (2002). Bayesian treed model. *Machine Learning*, **48**, 299–320.

DENISON, D. G. T. and MALLICK, B. K. (2000). Classification trees. *Biostatistics*, **5**, 365–372.

DUDA, R. O. and HART, P. E. (1973). *Pattern classification and scene analysis.* Wiley-Interscience, New York.

DUDA, R. O., HART, P. E. and STORK, D. G. (2001). *Pattern classification.* 2nd ed. Wiley-Interscience, New York.

EDWARDS, A. and CAVALLI-SFORZA, L. (1965). A method for cluster analysis. *Biometrics*, **21**, 362–375.

EVERITT, B. S. (1993). *Cluster analysis.* 3rd ed. Edward Arnold, London.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The elements of statistical learning.* Springer Series in Statistics, Springer-Verlag, New York. Data mining, inference, and prediction.

LORR, M. (1983). *Cluster analysis for social scientists.* 2nd ed. Jossey-Bass Publishers, San Francisco.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Baerkeley Symposium on Mathematical Statistics and Probability*, **1**, 281–297.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized linear models.* Chapman and Hall, London.

ROBERT, C. P. (2001). *The Bayesian choice.* 2nd ed. Springer Texts in Statistics, Springer-Verlag, New York.

SARLE, W. (1983). Cubic clustering criterion. Technical Report A-108, SAS Institute Inc., Cary, NC.

WARD, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.