STATISTICAL PREDICTION OF CANCELLATION BEHAVIOR AMONG HOLDERS OF MOTOR INSURANCE CONTRACTS IN GERMANY

A Thesis

Presented to the

Faculty of

San Diego State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Statistics

---

by

Andreas L. Reuss

Summer 2002

THE UNDERSIGNED FACULTY COMMITTEE APPROVES

THE THESIS OF ANDREAS L. REUSS

Duane L. Steffey, Chair Department of Mathematics and Statistics	Date

Chii-Dean Lin
Department of Mathematics and Statistics

Ming Ji
Graduate School of Public Health

SAN DIEGO STATE UNIVERSITY

Summer 2002

DEDICATION

To my family.

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Duane Steffey for his guidance and support throughout my work on this thesis. I also wish to thank Dr. Chii-Dean Lin and Dr. Ming Ji for serving on my thesis committee.

In addition, I would like to thank Prof. Hans-Joachim Zwiesler and Michael Pecheim from the University of Ulm, Germany, who provided the dataset and a significant amount of background information.

TABLE OF CONTENTS

PAGE

CHAPTER PAGE

CHAPTER PAGE

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

MOTOR INSURANCE IN GERMANY

The primary goal of this thesis is to investigate how statistical models can be used to predict the cancellation behavior among holders of motor insurance contracts in Germany. Before the underlying study and the provided dataset are discussed in detail, I want to give some necessary background information about motor insurance in Germany.

In Germany, the owner of a motor vehicle is required by law to compensate for damages caused by the operation of his motor vehicle (§7 (1) StVG[1]). This includes compensation for personal injuries or death and for damages to property. By §1 PflVG[2], the owner of a motor vehicle is required to conclude a contract with an insurance company that covers these types of damages, no matter whether they are caused by the owner himself or by another driver (motor third-party liability).

The counterpart of the obligation of the owner to purchase liability insurance for his motor vehicle is the requirement for insurance companies to accept any request for this type of liability coverage (§5 (2) PflVG). However, the insurance company can add a risk surcharge to the premium if it believes the contract represents a higher risk.

In addition to the required liability insurance, owners of motor vehicles can also purchase vehicle insurance and passenger accident insurance. The two latter types of

[1]Road Traffic Act (Strassenverkehrsgesetz)

[2]Compulsory Insurance Law (Pflichtversicherungsgesetz)

motor insurance are voluntary for the owner. There is also no obligation for the insurance company to grant coverage. Usually the different types of motor insurance are combined in a single contract. For more details see Asmus and Sonnenburg (1998, p. 51).

1.1 Types of Motor Insurance

Asmus and Sonnenburg (1998, chap. 5) discuss the different types of motor insurance. General conditions for motor insurance can be found in the AKB[3], which are usually very similar from one company to another. I will refer to the AKB used by the Versicherungskammer Bayern (VKB) after 7/1/1998 for illustration. The AKB define three types of motor insurance:

1. third-party liability (TPL) insurance,
2. partial or full motor vehicle own damage insurance and
3. passenger accident insurance.

Coverage in motor TPL insurance includes the satisfaction of justified demands and the defense against groundless demands for compensation against insured persons that are filed because

1. persons have been injured or killed (minimum coverage: 2.5 million Euro) or
2. property has been damaged or destroyed (minimum coverage: 500,000 Euro) or
3. basic rights of other persons have been violated without any damage to property (minimum coverage: 50,000 Euro)

by the operation of the insured motor vehicle (§10 (1) AKB). Both the owner and the actual driver are insured. The sums given are minimal legal requirements since 1997.

[3]General policy conditions for motor insurance (Allgemeine Bedingungen fuer die Kraftfahrtversicherung)

Motor vehicle own damage insurance provides coverage against damage, destruction and total loss of the insured automobile. Partial vehicle own damage insurance (semi-comprehensive cover) covers damages that are caused by fire, explosion, theft, thunderstorm, lightning or collision with wild animals. Full motor vehicle own damage insurance (comprehensive cover) additionally covers damages of the insured vehicle that are due to an self-inflicted accident. It also provides compensation for damages caused in a hit-and-run accident or by unknown persons (§12 (1) AKB). There may be deductibles for both the semi-comprehensive and the comprehensive cover.

Passenger accident insurance is also related to a specific vehicle. It covers damages caused by accidents of insured persons while operating the vehicle, repairing the vehicle and during related activities (§18 AKB).

1.2 Tariffs in Motor Insurance

The next step is to understand how tariffs in motor insurance are designed and how premiums are calculated. For this section I will closely follow the discussion in Johannson (1999, chap. 2).

In order to receive the benefits described in the previous section, the policy holder has to pay a premium to the insurance company. In general, the gross premium is calculated according to the following scheme:

	net risk premium
+	security surcharge
=	net premium
+	overhead surcharge
+	profit surcharge
=	gross premium

The equivalence principle states that the net risk premium must be equal to the expected value of future payments for claims in one insurance period. In case of motor insurance, the insurance period is usually one year. The net risk premium is supposed to cover only the pure compensation costs. A security surcharge is added to account for the risk that the actual payments for claims exceed the expected value. The overhead surcharge covers all other costs, including commissions and general administration costs. Finally, a profit surcharge may be added.

Since the distribution of future compensation payments is unknown, we have to estimate its expectation in order to determine the net risk premium. The variance of this distribution may also be needed to determine the security surcharge.

In order to do this, contracts with similar risks are put into groups that are as homogeneous as possible (risk collective). This is achieved by using characteristics that influence the number and size of future claims (risk factors). We distinguish between objective risk factors (related to objects) and subjective risk factors (related to persons).

Then the risk factors with the strongest influence on the total claim size (tariff factors) are identified, often using statistical methods. On the basis of these tariff factors, tariff classes are set up and the appropriate net premium for every class is determined (tariff calculation).

Major changes in the tariff calculation occured when motor insurance was deregulated in 1994. Hence tariff design and premium calculation before and after the deregulation will be discussed separately.

1.2.1 Before the Deregulation

Until the deregulation in 1994, the design and the calculation of the tariffs in motor TPL insurance was subject to strict legal regulations. A federal office, the BAV[4], checked the compliance of the tariffs with these regulations. Any change in the design, calculation and application of a tariff had to be preapproved by the BAV.

Furthermore, all insurance companies were required to give precisely specified information on the number and size of claims, and the tariff factors for all contracts in their portfolio. These data were then combined to a general claims statistic. After smoothing the data using statistical methods, the resulting calculation statistic was the basis for all tariffs and premiums of all insurance companies in Germany. This means that the net premiums were basically administered by the BAV.

In 1993, three tariff factors were used to set up the calculation statistic: vehicle power, regional class and no-claims class. The characteristics vehicle power and regional class were used to determine a basic premium. Then a discount or surcharge was applied to the basic premium according to the no-claims class to get the net premium.

These strict regulations did not apply for motor vehicle own damage insurance (semi-comprehensive and comprehensive cover) and for passenger accident insurance. These types of motor insurance had been deregulated before, in 1985, 1982 and 1979 respectively (see Asmus and Sonnenburg (1998, p. 24)).

I will now take a closer look at the tariff factors for motor TPL because they will appear as variables in the dataset.

[4]Federal Insurance Supervisory Office (Bundesaufsichtsamt fuer das Versicherungswesen)

1.2.1.1 Vehicle power. The engine power, measured in kilowatts (kW) was the only tariff characteristic that was directly related to the insured vehicle. It is reasonable to believe that "bigger" vehicles are likely to produce "bigger" claims. In 1993, vehicles were classified into 11 classes according to engine power.

1.2.1.2 Regional class. The tariff factor regional class was a combined characteristic of profession and regional classification. Three different groups of professions were distinguished: agricultural professions (A), public employees (B) and others (R). Group B was then divided into six regional classes (B1 to B6), whereas Group R was divided into five rural classes (RL1 to RL5) and five urban classes (RS1 to RS5), according to the expected total claim size. Since group A made up only 5% of the total portfolio, there was no regional classification for this group.

1.2.1.3 No-claims class. The tariff characteristic no-claims classs took into account the individual claims history of a specific contract. Insured persons that had not caused any claim in the past years were classified into higher no-claims classes depending on the number of years without a claim. All other contracts were classified into lower no-claims classes. For each class a percentage rate was defined that determined the percentage of the basic premium that actually had to be paid by the policy holder. There were different no-claims classes for motor TPL, semi-comprehensive and comprehensive vehicle insurance.

1.2.2 After the Deregulation

On 7/29/1994, the motor TPL insurance in Germany was deregulated by a law that set equal standards for the insurance industry throughout the European Union. The core of the deregulation was the abolition of the obligation to use the same tariffs. From this time on, new tariffs do not have to be approved by the BAV.

The design of the tariff and the calculation of the premium is now up to the insurance company. It can charge the premium that is necessary according to its own opinion. It can also modify the design of the tariff, for example by modifying existing tariff factors or by introducing new tariff factors that seem useful. Many insurance companies made use of these possibilities in the years after 1994. But the tariffs of different companies in Germany still show a lot of similarities. I will use the tariff provisions (TB) used by the VKB after 7/1/1998 for illustration, since they apply to the newest contracts in the provided dataset.

Overall, the most important modification was the replacement of the tariff factor "vehicle power" with a new characteristic "type class" by all companies in 1996. An empirical investigation had shown that this characteristic was more suitable to predict the number and size of future claims than the old characteristic "vehicle power".

Furthermore regional classification and profession are now separate tariff factors. As most companies, also the VKB introduced new tariff factors after 1996, for example mileage per year, age of the vehicle, number of drivers and a home owner characteristic (§6 TB). All these tariff factors will appear in the dataset and therefore need to be discussed in more detail.

1.2.2.1 Type class (§12 TB). Since 1996, most companies use the vehicle type as one of the factors to determine the basic premium. For every available type of vehicle, the ratio of the average total claim size caused by this vehicle to the average total claim size caused by all vehicles is determined. This index is then used to form groups of vehicles that should have similar expected future claim sizes. There are different type classes for motor TPL, partial motor vehicle own damage and full motor vehicle own damage insurance. An example for motor TPL insurance is given in Table 1.1.

Table 1.1: Type classes for motor TPL (from §12 (3) TB)

Type class	Index for the average total claim size of the vehicle
10	below 49.5
11	[49.5,61.9)
12	[61.9,71.6)
13	[71.6,79.8)
14	[79.8,86.6)
15	[86.6,92.0)
16	[92.0,97.7)
17	[97.7,103.7)
18	[103.7,110.4)
19	[110.4,118.0)
20	[118.0,125.4)
21	[125.4,133.3)
22	[133.3,144.0)
23	[144.0,165.4)
24	[165.4,196.0)
25	over 196.0

1.2.2.2 Regional class (§8 and §11 TB). Similar to the type class, for every one of the 446 registration districts in Germany, the ratio of the average total claim size in this district to the average total claim size over all districts is determined. The resulting index is then used to form several regional classes. There are different regional

classes for motor TPL, semi-comprehensive and comprehensive vehicle insurance. An example for motor TPL insurance is given in Table 1.2.

Table 1.2: Regional classes for motor TPL (from §8 (2) TB)

Regional class	Index for the average total claim size in the registration district
1	below 84.7
2	[84.7,90.7)
3	[90.7,93.6)
4	[93.6,95.8)
5	[95.8,98.3)
6	[98.3,100.8)
7	[100.8,103.9)
8	[103.9,106.9)
9	[106.9,111.1)
10	[111.1,115.4)
11	[115.4,120.0)
12	over 120

1.2.2.3 No-claims class (§14 - §19 TB). At the beginning of every insurance period, each contract is classified into a no-claims class. If the contract has lasted at least one year without a claim, then the contract is classified into a no-claims class for the following year, depending on the number of years without a claim. For example, a contract with x years without a claim is classified in class SF x for $1 \leq x \leq 25$. Contracts that have been without claim more than 25 years are also in class SF 25. There is an additional no-claims class SF 1/2. It applies under any of the following conditions:

- If the contract has lasted less then one year without a claim.
- If the policy holder is classified into a no-claims class for his first vehicle and effects a motor TPL insurance policy for a second vehicle.
- If a contract for a new vehicle is established and both the spouse of the policy holder is classified into a no-claims class and the policy holder has had a driving licence for a least one year.
- If the policy holder has had a driving licence for at least three years and effects a new policy.

There are also two claims classes for motor TPL insurance, denoted by S and M. If a contract cannot be classified in any of the classes S, M, SF 1/2 and SF 1 to SF 25, then it is classified into class 0.

If one or more claims occur within one insurance period, the classification of the contract into a no-claims class changes in the following year. An example of rules for motor TPL insurance is given in Table 1.3.

Table 1.3: Change of the no-claims class (from §19 (1) TB)

	1 claim	2 claims	3 claims	4 or more claims
From class	Into class			
SF 25	SF 18	SF 8	SF 3	M
SF 18 - SF 24	SF 10	SF 4	SF 1	M
SF 17	SF 7	SF 3	SF 1	M
SF 16	SF 7	SF 3	SF 1	M
SF 15	SF 6	SF 2	SF 1/2	M
SF 14	SF 6	SF 2	SF 1/2	M
SF 13	SF 5	SF 2	SF 1/2	M
SF 12	SF 5	SF 2	SF 1/2	M
SF 11	SF 5	SF 2	SF 1/2	M
SF 10	SF 4	SF 1	SF 1/2	M
SF 9	SF 4	SF 1	SF 1/2	M
SF 8	SF 4	SF 1	SF 1/2	M
SF 7	SF 3	SF 1	SF 1/2	M
SF 6	SF 2	SF 1/2	S	M
SF 5	SF 2	SF 1/2	S	M
SF 4	SF 1	S	M	M
SF 3	SF 1	S	M	M
SF 2	SF 1/2	0	M	M
SF 1	SF 1/2	0	M	M
SF 1/2	S	M	M	M
S	M	M	M	M
0	M	M	M	M
M	M	M	M	M

For each class a corresponding percentage rate is defined that applies as a discount or surcharge on the basic premium. An example is given in Table 1.4 .

Table 1.4: Percentage rates for the no-claims classes (from §18 (1) TB)

No-claims class	Percentage rate TPL insurance	Percentage rate vehicle insurance
SF 18 - SF 25	30	30
SF 17	35	35
SF 16	35	35
SF 15	35	35
SF 14	40	35
SF 13	40	40
SF 12	40	40
SF 11	40	45
SF 10	45	45
SF 9	45	45
SF 8	50	50
SF 7	50	55
SF 6	55	60
SF 5	60	65
SF 4	65	70
SF 3	75	80
SF 2	85	90
SF 1	100	100
SF 1/2	120	115
S	155	-
0	240	190
M	245	-

1.2.2.4 Tariff group (§9 and §10 TB). The contract is classified into one of the following main tariff groups: 'A' for farmers, 'B' for public employees and public corporations and 'N' for all others. The tariff group is another factor that affects the basic premium.

1.2.2.5 Mileage class (§13 TB). According to the mileage (per year) of the insured vehicle, a contract is classified into one of the following mileage classes with corresponding discounts or surcharges on the basic premium for motor TPL and motor vehicle insurance (see Versicherungskammer Bayern [VKB] (2001, p. 4)):

- 1 ($< 9,000$ km): - 15%
- 2 ($9,000 - 12,000$ km): - 10%
- 3 ($12,000 - 17,000$ km): 0%
- 4 ($17,000 - 30,000$ km): + 5%
- 5 ($> 30,000$ km): + 10%

The insurance company has the right to verify the actual mileage.

1.2.2.6 Age of vehicle (§4 (1) TB). The age of the vehicle at the time of the purchase by the policy holder can also result in a discount or surcharge on the basic premium. For example, in motor TPL the discounts are structured as follows (see VKB (2001, p. 4)):

- not older than 1 year: - 15%
- not older than 4 years: - 10%
- not older than 7 years: 0%
- older than 7 years: + 20%

1.2.2.7 Number of drivers (§13a TB). If the vehicle is only used by the policy holder and his spouse, an additional discount may apply, depending on the age of the vehicle. Additional restrictions are that both policy holder and spouse must be at least 25 years old and that the contract must be classified into a no-claims class. For details see VKB (2001, p. 4).

1.2.2.8 Home owner characteristic (§13b TB). If the policy holder owns the house he is currently living in and if he has a fire insurance contract with the same insurance company, then he qualifies for another discount of 10% (see VKB (2001, p. 4)).

1.3 Economic Importance of Motor Insurance

In the last section of this introductory chapter I want to emphasize the economic importance of motor insurance in Germany, and also motivate the cancellation prophylaxis study. Together with life insurance and private health insurance, motor insurance is one of the key sectors of the German insurance industry. I want to provide some information about this sector which can be found in the 2001 yearbook of the German Insurance Association[5].

In 2000, approximately 15.5% (20.36 billion Euro) of the gross premium income in the private insurance business in Germany were charged for motor insurance contracts (see Table 1.5 from GDV (2001, p. 51)).

If we consider the total number of insurance contracts, the percentage share of the motor insurance is even higher. At the end of 2000, 24.8% (97.21 million) of all insurance contracts were motor insurance contracts, followed by 87.49 million (22.3%) life insurance contracts (see Table 1.6 from GDV (2001, p. 62)).

In 2000, more than 81% of all households had a motor TPL insurance contract and 33% had a full motor vehicle own damage insurance (see Table 1.7 from GDV (2001, p. 59)). Currently there are approximately 51 million vehicles on the

[5]Gesamtverband der deutschen Versicherungswirtschaft (GDV)

Table 1.5: Premium income in private insurance (2000)

Insurance class	Gross premiums written (in Euro bn)
Life insurance	60.95
Private health insurance	20.71
Motor insurance	20.36
Property insurance	12.30
General liability insurance	5.87
Private accident insurance	5.40
Legal expenses insurance	2.69
Marine insurance	1.64
Credit, aviation, nuclear insurance	1.52
Others	0.15
TOTAL	131.59

Table 1.6: Insurance portfolios as at year-end (2000)

Insurance class	Insurance contracts (in millions)
Motor insurance	97.21
Life insurance	87.49
Property insurance	67.11
Private health insurance	47.85
General liability insurance	37.81
Private accident insurance	29.06
Legal expenses insurance	28.72
TOTAL	392.25

road (registered). This explains, together with the obligation to purchase motor TPL insurance, the large number of contracts in motor TPL insurance.

All these numbers show that the motor insurance is one of the key sectors of the German insurance industry and that it has considerable overall economic importance. Details about the premium income and the claims expenditure in motor insurance in Table 1.8 (from GDV (2001, p. 82)) show that in 2000 income and expenditure met at 20.36 billion Euro. For that year the loss ratio (the share of gross claims expenditure on claims of the financial year in premiums earned in percent) was roughly 100.

Table 1.7: Insurance cover of households (2000/2001)

Insurance class	Households with insurance contract (in %)
Motor TPL insurance	81.2
Comprehensive insurance on contents	77.2
Private liability insurance	65.2
Life insurance	54.6
Legal expenses insurance	43.2
Private accident insurance	40.1
Motor vehicle full own damage insurance	33.3
Private health insurance only	11.8

Table 1.8: Premium income in motor insurance (2000)

Insurance class	Gross premiums written (in Euro bn)
Motor TPL	12.63
Full own damage cover	5.75
Partial own damage cover	1.74
Passenger accident	0.24
TOTAL gross premium income	20.36
TOTAL gross claims expenditure	20.36

Since after the deregulation in 1994 insurance companies offering motor TPL insurance are allowed to modify their tariffs and to introduce new tariff factors, there is now a wide variety of tariffs for motor TPL insurance. Customers are more than ever before aware of their choices and realize that they may save a considerable amount of money by cancelling their current contract and effecting a new policy with another insurance company that offers a lower premium. Overall, competition between insurance companies has significantly intensified.

In such a price-sensitive market it is important for the companies to find ways of preventing their customers from cancelling. In general, it is much more expensive for the insurance company to acquire a new customer than to prevent a current customer

from cancelling. Recent general studies estimate that, on average, it is five times more expensive to acquire a new customer than to maintain relations with a current customer (see Wirtz (2000, p. 160)). In order to acquire new customers, the company needs to offer premiums that are lower than the premiums of the major competitors. Of course, there are other important aspects like the quality of customer service and distribution system, but the premium is the most important factor in this market.

On the other hand, people who already have a motor insurance contract with some company are usually less sensitive to price differences. They tend to keep their contract even if another company would offer them a lower premium. In order to provide the insurance company with information on the cancellation behavior of their customers, the cancellation prophylaxis study discussed in the chapter 2 was designed.

CHAPTER 2

THE CANCELLATION PROPHYLAXIS STUDY

In this chapter I will introduce the cancellation prophylaxis study and the dataset that was provided for the analysis.

2.1 Cancellation by the Policy Holder

Before actually discussing the cancellation prophylaxis study in detail, I want to explain under which circumstances the policy holder has the right to cancel the motor insurance contract and which deadlines apply for cancellation. For a detailed discussion see Asmus and Sonnenburg (1998, p. 133).

First it should be noted that in case of cancellation of the contract, the policy holder usually does not lose the possibility of getting a discount according to the number of years without claim (no-claims class). §5 (7) PflVG and §25 TB state that after the cancellation of a contract, the insurance company must issue a certificate containing information about the length of the contract, the number of years without a claim and, if there were any, the number of claims covered by the policy. If the policy holder concludes a new motor insurance contract with another insurance company, he is usually classified into a no-claims class corresponding to the certified number of years without claim (§24 TB).

If, on the other hand, the old contract was classified into one of the claims classes S or M, the policy holder may try not to inform the new insurance company about this fact in order not to be classified into one of these classes but in the class SF 1/2. However, the new insurance company can demand the certificate from the old insurance company and it will do so in general (§25 TB).

In general, there are two types of termination of a contract: ordinary and extraordinary termination. Both types will be discussed with respect to the right of the policy holder to terminate the contract.

2.1.1 Ordinary Termination

Conditions for ordinary termination can be found in §4 AKB. The duration of a motor insurance contract can be one year or less. If the duration is less than one year, then the contract ends automatically at the expiration date. For one-year contracts, the contract automatically extends for another year if it is not cancelled at least one month before the expiration date. The cancellation notice must be in written form. If the policy holder cancels, then the contracts ends at the end of the insurance period. Termination may refer to the complete contract or only to parts, for example only to vehicle insurance.

2.1.2 Extraordinary Termination

There are several possible reasons for extraordinary termination of the contract. Different reasons may lead to the cancellation, the deadlines may differ and the contract may not necessarily end at the end of the insurance period.

2.1.2.1 Cancellation in case of a claim (§4b AKB). In case of occurrence of one of the events insured against, the policy holder has the right to cancel the contract. This can be done as soon as the insurance company has either accepted or refused the demand for compensation or as soon as the insurer has advised the policy holder to take legal action against the demand for compensation, but no later than one month after the policy holder received notice about this fact. The contract may end immediately or at any time until the end of the current insurance period. There is no refund for the premiums paid for the current insurance period.

2.1.2.2 Cancellation after changes of the tariff (§9a, 9b, 9c, 9d AKB). Changes of the underlying tariff may affect the premium a policy holder has to pay for his motor insurance contract. This may happen in three different ways:

1. The rules for the application of tariff factors as described in section 1.2.2 may be modified by the insurance company. Modifications then apply at the beginning of the next insurance period (§6 TB).
2. Every year an independent trustee uses statistics from a sufficiently large number of insurance companies to find the appropriate index values that are used to determine the type class of a vehicle and the regional class of a registration district. Then the classification of a specific contract into a regional class or into a type class may change at the beginning of the next insurance period (§11 and §12 TB).
3. Changes in legal requirements may also affect the tariff, for example by the obligation to include new types of coverage or to grant higher maximum compensation sums (§9c AKB). There may also be changes in the conditions of the contract (§9d AKB).

§9a (1) and §9c (1) AKB state that each of the three cases above may lead to a premium increase by the insurance company beginning with the next insurance period (case 1 and 2) or as soon as the changes apply (case 3). The policy holder must receive a written notice about the change of the premium at least one month before it applies

(§9a (2) AKB). Then he has the right to cancel the contract within one month (§9a AKB). The contract ends at the time the premium increase would apply. In case 1, the policy holder can cancel even if the changes do not increase his premium.

§9a (3) AKB states that the policy holder has no right to cancel if the premium increases due to classification into a different regional class, tariff group or type class for which the policy holder is responsible (for example, if the policy holder moves into another registration district, changes his profession or if there were claims covered by the contract).

2.1.2.3 Cancellation after the sale of the vehicle (§6 AKB). If the policy holder sells the insured vehicle, the buyer takes over all rights and responsibilities of the contract. Both buyer and seller are liable for the premium of the current insurance period.

Within one month after the purchase or the notice about the existence of an insurance contract, the buyer has the right to cancel the contract. The contract may end immediately or at any time until the end of the current insurance period. Only the premium for the actual time of coverage will be charged.

The seller, however, has the right to be classified into the same no-claims class as before if he concludes a motor insurance contract for a new vehicle. Certain restrictions apply (§23 TB).

2.1.2.4 Cancellation after the death of the policy holder. If the policy holder dies, the heirs take over all rights and responsibilities of the contract. Within one month after taking over the vehicle, the heirs have the right to cancel the contract.

The contract may end immediately or at any time until the end of the current insurance period. Only the premium for the actual time of coverage will be charged (see Asmus and Sonnenburg (1998, p. 139)).

2.2 Description of the Problem

We have seen in the previous section that there are several circumstances under which the policy holder has the right to cancel his motor insurance contract. As discussed in section 1.3, the insurance company generally wants to prevent its customers from cancelling their contracts. Even if a contract has caused many claims in the past, the insurance company usually does not want the policy holder to cancel because the no-claims class system automatically leads to a premium increase for this customer (see percentage rates for the claims-classes S and M in Table 1.4).

Now the insurance company is interested in knowing in advance which contracts are likely to be cancelled by the policy holder. If it knew in advance that an individual policy holder is considering to cancel his contract, the insurance company could try to prevent him from cancelling (cancellation prophylaxis). For example, the insurance agent may be informed about contracts that are likely to be cancelled. The agent then tries schedule a meeting with the policy holder to discuss possible changes in his contract. This includes changing into the newest tariff generation with possible premium reductions and checking whether additional discounts (home owner, limited mileage etc.) may be applicable to the contract.

In this thesis I am not discussing possible actions that may be taken to prevent customers from cancelling their contracts. Instead I am concentrating on identifying risk factors for cancellation and on setting up a classification rule that discriminates between contracts that are likely to be cancelled and contracts that are not. Therefore the setup for the study is the following:

1. Randomly choose contracts from the current total portfolio of the insurance company and contracts that have been cancelled recently. Store all available characteristics that describe the sampled contracts.
2. Use the available characteristics of the contracts in the sample to set up a statistical model that describes the dependence of the cancellation behavior on these characteristics and provides a rule for classifying contracts as likely to be cancelled or unlikely to be cancelled for every possible combination of the characteristics. A reasonable time frame for cancellation would be one insurance period (one year).
3. Finally use the classification rule to identify contracts in the total portfolio that are likely to be cancelled. For these contracts cancellation prophylaxis actions should be considered.

Although the primary goal is to find an accurate classification rule, the company also wants to understand which variables influence the cancellation behavior and quantify their influence.

2.3 Study Design

The cancellation prophylaxis study is a typical case-control study. We "look in the past" and observe the outcome on several subjects (retrospective design). For a general description of case-control studies see for example Hosmer and Lemeshow (2000, p. 206).

More specifically, the population of all contracts is first divided into two strata by the binary outcome "cancellation". One strata contains all contracts cancelled between 1996 and 1998 (cases) and the other strata contains all contracts that are still in the

portfolio in 1999 (controls). Then samples of fixed size (in this case, 10,000) are chosen from the two strata and all available characteristics (in this case, 40) are measured for each contract sampled. We assume that all relvant characteristics are available in the computer system. This sample of 20,000 contracts should then be used to analyze the dependence of the outcome cancellation on the characteristics of the contract.

One consequence of this design is that we cannot directly estimate the proportion of contracts that are cancelled. In order to do so, we need to know the total number of cancelled and not cancelled contracts. We also cannot estimate the actual cancellation probability for a given contract in the portfolio; the sampling design only allows to classify contracts as likely or unlikely to be cancelled.

An alternative sampling design would be a cohort study. In this prospective design, we would chose a random sample from the total portfolio of motor insurance contracts at some time and determine the values of the covariates. Then we would follow the chosen contracts for a fixed period of time (say, for example, one year). At the end of that time period we would check whether the contract has been cancelled or not. Under this design we would be able to estimate the probability of cancellation for every contract in the portfolio.

2.4 Description of the Dataset

The given dataset is provided by the Versicherungskammer Bayern (VKB), one of the largest motor insurance companies in Germany. According to the annual report for 1999, the total portfolio of motor insurance contracts of this company consisted of

940,500 motor TPL insurance policies out of which 678,491 included motor vehicle own damage insurance. Gross premiums amounted to 209.0 million Euro for motor TPL insurance and 121.4 million Euro for motor vehicle own damage and passenger accident insurance .

A brief description of each of the 40 variables in the dataset is given in appendix A. As discussed there, 28 variables seem generally useful for this study. One of them (STORNO) represents the binary outcome of interest (contract cancelled or not).

Several variables are related to the policy holder: age (KUNALTER2), nationality (AUSLK1), zip code (PLZ), county (REGBE), payment option (ZART) and the number of years the person had a motor insurance contract with the VKB (VWDAU). Useful information about the insured vehicle is available through the variables STAERKE (vehicle power in kW), FALTER (age) and KAUFDAT1 (year of purchase).

As far as the motor TPL component of the insurance contract is concerned, we have information on tariff generation (KZKHT), type of coverage (DECKAR), regional classification (REGIO), tariff group (TGR1), restricted mileage (KZWF), restricted number of drivers (FANZ1), home owner discount (KZEHB1), type class (TYPKLH), no-claims class (TARKLA), percentage rate (BEITSAT) and annual premium (JTBH).

The variable KZFV indicates whether motor vehicle own damage insurance is included in the contract. If this is the case, then several variables describe this part of the contract, namely TYPKL (type class), FELD66 (no-claims class), FELD68 (percentage rate) and JTBF (annual premium). Finally, additional discounts are coded in the variables SORAB1 (special discount) and VERBU1 (employee tariff).

Among those 27 variables 9 are measured on interval or ratio scale (VWDAU, KUNALTER2, STAERKE, FALTER, KAUFDAT1, BEITSAT, JTBH, FELD68, JTBF), 3 are measured on ordinal scale (ZART, TYPKLH, TYPKL) and 15 on nominal scale (AUSKL1, PLZ, REGBE, KZKHT, DECKAR, REGIO, TGR1, KZWF, FANZ1, KZEBH1, TARKLA, KZFV, FELD66, SORAB1, VERBU1).

2.5 General Modeling Considerations

The general setting for this study is the following: The response (or outcome or dependent) variable of interest is binary (dichotomous):

$$y_i := \begin{cases} 1 & \text{if the i-th contract is cancelled within the next insurance period} \\ 0 & \text{if the i-th contract is not cancelled} \end{cases} .$$

For every motor insurance contract in the sample we have observed whether it has been cancelled or not. Additionally, several characteristics stored in the computer system were recorded for all 20,000 contracts. I follow the usual terminology and combine these variables into a vector of explanatory (or predictor or independent) variables $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})$. According to section 2.4, all types of measurement scales (nominal, ordinal, interval and ratio) occur in the given dataset and there seem to be $p = 27$ useful explanatory variables. Any model under consideration should be able to handle this set of explanatory variables.

For the purpose of setting up a statistical predictive model that describes the dependence of the outcome y_i on the explanatory variables $\mathbf{x_i}$, a common assumption is that the outcomes for different subjects are independent. In this context, it means that the cancellation behavior of a specific contract does not depend on the cancellation

behavior of other contracts in the portfolio. There may be situations where this is not true, for example, if a group of people receives an advertising message from a competitor and subsequently cancels. However, since the underlying population is large (the portfolio of the insurance company consists of approximately 1 million contracts) and contracts are selected randomly, the assumption seems justifiable. Thus, the outcomes y_i are treated as independent binary random variables. If there are reasons to believe that outcomes are correlated, then some of the standard models for categorical data analysis can be adjusted (see Agresti (1990, p. 456)).

The resulting model is required to have the following properties (which are common requirements for statistical models):

- The resulting model should fit the data, while being as parsimonious as possible.
- It should also be reasonable and interpretable in the context of the study.
- And, of course, it should be able to predict the outcome as accurately as possible.

In this thesis, I will investigate the use of two types of models for the prediction of cancellation behavior: logistic regression models and classification trees. As Hosmer and Lemeshow (2000) note, logistic regression models have become the standard method of analysis in a situation where the goal is to describe the relationship between a discrete (often binary) outcome and several explanatory variables. Logistic regression models can be seen as an adaptation of the standard regression model to the case of a discrete dependent variable, within the common framework of Generalized Linear Models (see Agresti (1990)).

An alternative approach is the use of trees for prediction. The basic idea of a classification tree is to recursively partition the covariate space into more and more

homogeneous subsets. The partition is achieved by binary recursive splitting and can be described graphically by a tree. Classification trees are a computer-intensive technique and have a wide range of applications. They are especially useful if the number of observations and the number of covariates is large.

I will discuss logistic regression models and classification trees in more detail in chapters 3 and 4, respectively. Hastie, Tibshirani, and Friedman (2001) give a comprehensive overview of statistical predictive models.

CHAPTER 3

LOGISTIC REGRESSION MODELS

This chapter summarizes the basic methodology for logistic regression models, as well as some aspects of model selection and model assessment.

3.1 General Idea

In a logistic regression model we assume that the binary outcome Y_i can be modeled as a random variable having a Bernoulli distribution with some unknown probability of success that depends on the vector of covariates $\mathbf{x}_i$:

$$Y_i \sim \text{Bernoulli}(\pi(\mathbf{x}_i)), \ 0 \leq \pi(\mathbf{x}_i) \leq 1.$$

We first observe that

$$\text{E}(Y_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i) \quad \text{and} \quad \text{Var}(Y_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i)).$$

The conditional expectation of Y_i given $\mathbf{x}_i$ must lie in the interval $[0, 1]$, and the conditional variance depends on $\mathbf{x}_i$. The classical linear regression model does not satisfy these conditions and must therefore be modified in order to be used for a binary response variable.

The basic idea is to consider the logit transformation of the probability of success $\pi(\mathbf{x}_i)$ which maps the range of probabilities $(0, 1)$ onto the interval $(-\infty, +\infty)$ and is

defined by:

$$\text{logit}\left(\pi(\mathbf{x}_i)\right) := \ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right). \tag{3.1}$$

Then a linear predictor is used in the same way as for the standard linear regression model. This gives the following model equation for the case of p explanatory variables:

$$\ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}. \tag{3.2}$$

Equivalently we have

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}},$$

where the right-hand side can be recognized as the cumulative distribution function (c.d.f.) of the logistic distribution. The question is why exactly this type of distribution (this particular transformation) is chosen.

First of all, the choice of c.d.f. of the logistic distribution is motivated by plots of the proportions of positive responses versus the values of continuous covariates. These curves often have an S-shaped form that looks like the c.d.f. of a continuous probability distribution. The c.d.f. of the logistic distribution in particular has a simple form which can be easily manipulated and corresponds to a rather simple transformation, the logit. Using the c.d.f. of the standard normal distribution instead (as in probit models) results in a transformation that cannot be expressed explicitely.

Second, it leads to a meaningful interpretation of the model in terms of odds ratios. For details see section 3.6. Third, logistic regression models can be used for the analysis of data from case-control studies in the same way as for data that were collected in a cohort study (Hosmer and Lemeshow (2000, p. 208)). Finally, it is also theoretically

appealing, because the logit link is the canonical link for Generalized Linear Models with Bernoulli random component (see Agresti (1990, p. 80)).

Both continuous and categorical variables can be used as predictors in logistic regression models as given by Equation (3.2). For a categorical variable with k levels we simply define $k-1$ dummy variables and use them in the formulation of the model.

Multiple logistic regression models as introduced above are analogous to the normal theory linear models and have become the standard models for data analysis in situations comparable to this study. A more detailed general discussion of logistic regression models can be found in Agresti (1990), Cox and Snell (1989), and Hosmer and Lemeshow (2000).

3.2 Fitting Logistic Regression Models

After settting up a logistic regression model, the first step of the analysis is to estimate the unknown parameters $\beta_i,\ i = 0, \ldots, p$. All major software packages use the maximum likelihood method, although there are alternative approaches (see Hosmer and Lemeshow (2000, p. 21)).

For logistic regression models the likelihood equations are nonlinear in the unknown parameters. They can be solved iteratively using the Newton-Rhapson algorithm; this algorithm also computes the estimated asymptotic covariance matrix of the parameters as a byproduct. The iterative process of calculating the maximum likelihood estimates is sometimes referred to as iterative weighted least squares algorithm. The parameter estimates can then be used to estimate the logit for a particular

observation and the probability of success by inverting the logit transformation:

$$\hat{\pi}_i := \hat{\pi}(\mathbf{x}_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_p x_{ip}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_p x_{ip}}}.$$

Likelihood equations and the Newton-Rhapson algorithm are discussed in more detail in Agresti (1990, p. 112) and in Hosmer and Lemeshow (2000, p. 33).

3.3 Significance Testing

After estimating the coefficients, the next step is to assess the significance of the predictor variables in the model. The natural question is whether a model M_1 that includes some explanatory variables under consideration gives more information about the outcome than a model M_2 that does not include these variables. This question can be answered by testing the null hypothesis that M_2 holds, given that M_1 holds.

Let β^{M_1} and β^{M_2} be the parameter vectors for M_1 and M_2, respectively, such that β^{M_1} can be written in the form $\beta^{M_1} = \left(\beta^{M_2}, \gamma\right)$. That is, γ is a q-dimensional vector that contains the parameters for variables that are in M_1 but not in M_2. The test stated above is equivalent to considering M_1 and testing the null hypothesis $H_0 : \gamma = 0$ against the two-sided alternative $H_1 : \gamma \neq 0$. Note that in order to assess the significance of a categorical predictor with k levels, we have to test whether the coefficients of all $k-1$ dummy variables are equal to zero.

Three different test statistics have been suggested for this test. First, the Wald test is based on the asymptotic normality of the maximum likelihood estimates for M_1, denoted by $\hat{\beta}^{M_1} = \left(\hat{\beta}^{M_2}, \hat{\gamma}\right)$. The test statistic is given by

$$W = \hat{\gamma}'[\hat{C}ov(\hat{\gamma})]^{-1}\hat{\gamma}\ ,$$

where $\hat{Cov}(\hat{\gamma})$ denotes the estimated covariance matrix corresponding to the parameter estimates $\hat{\gamma}$. Second, the likelihood ratio test is based on the ratio of the maximized likelihoods for M_1 and M_2. The test statistic is

$$G = -2 \ln \left(\frac{\text{maximized likelihood for } M_2}{\text{maximized likelihood for } M_1} \right).$$

And third, the score test is based on the distribution of the partial derivatives of the log likelihood for M_1 with respect to the parameter vector β^{M_1}. Let $U(\beta^{M_1})$ be the vector of first partial derivatives of the log likelihood, and let $H(\beta^{M_1})$ be the matrix of second partial derivatives. In addition, let $I(\beta^{M_1})$ be either $-H(\beta^{M_1})$ or the expected value of $-H(\beta^{M_1})$. Now calculate the maximum likelihood estimates $\hat{\beta}^{M_2}$ for M_2 and set $\hat{\beta}^{(0)} = \left(\hat{\beta}^{M_2}, 0, \ldots, 0\right)$. The score statistic is defined by

$$S = U'(\hat{\beta}^{(0)}) I^{-1}(\hat{\beta}^{(0)}) U(\hat{\beta}^{(0)}).$$

Under H_0, all three test statistics have a large-sample chi-square distribution with q degrees of freedom. In the univariate case ($q = 1$), this fact can be used to find an approximate confidence interval for one of the unknown parameters β_j. For details see Cox and Snell (1989, p. 179).

Both Wald and likelihood ratio test require the computation of the maximum likelihood estimates for M_1. Further investigation has indicated that the Wald test is less powerful than the likelihood ratio test and that it can even show aberrant behavior: it may fail to reject the null hypothesis when the coefficient is significant. Therefore the likelihood ratio test is usually recommended (see Agresti (1990, p. 89)). The Score test, on the other hand, does not require the fitting of the larger model M_1. This reduced

computational effort is often cited as its major advantage, especially if several tests have to be performed in a stepwise procedure (see Hosmer and Lemeshow (2000, p. 16)).

3.4 Model Selection Procedures

The significance tests described in the previous section are the basis for stepwise procedures for selection or deletion of variables from a model. All these procedures check for the importance of a variable by measuring the statistical significance of the coefficient(s) of that variable. The statistical significance can be measured by computing p-values from any of the three test statistics mentioned in the previous section. Hosmer and Lemeshow (2000, chap. 4) discuss the commonly used stepwise procedures for model selection, namely backward elimination, stepwise selection (forward selection with backward elimination) and best subset selection.

These procedures are very useful and effective data analysis tools. However they only take into account the statistical significance of explanatory variables. If the sample size is large, effects that not are significant in the context of the problem may become statistically significant.

Furthermore, significance tests compare models in a relative sense; they do not consider whether the predicted values are an accurate representation of the observed values in an absolute sense. The latter is what is usually referred to as goodness of fit. Therefore stepwise procedures can give interesting insight in the structure of the problem, but further considerations as discussed in the following section are necessary in order to select an appropriate predictive model.

3.5 Assessing Goodness of Fit

In order to assess the goodness of fit of a logistic regression model we compare the observed outcomes $y_i \in \{0, 1\}$ to the fitted values $\hat{\pi}_i \in (0, 1)$ from the model. A general problem in this context is that the fitted model tends to perform in an optimistic manner on the dataset that was used for fitting. If the sample size is large enough (as it is the case for this study), a more unbiased assessment of goodness of fit can be achieved by using external validation: the original dataset is randomly divided into a training set and a validation set. The training set is used to fit the model; then the validation set is considered for the assessment of goodness of fit. Most authors recommend to split the data into 2/3 training and 1/3 validation set, or similar ratios. Hastie et al. (2001, chap. 7) discuss model selection and model assessment for statistical predictive models in general.

3.5.1 Pearson Chi-Square and Likelihood Ratio Statistic

Overall measures of the distance between y_i and $\hat{\pi}_i$ have been proposed for the assessment of goodness of fit. For this purpose it is useful to consider the data as described by a $J \times 2$ contingency table. The two columns are defined by the binary outcome y_i; the covariate patterns define the J rows. We will see that for every reasonable model in this study the number of covariate patterns is equal to the number of observations ($J = n$). From the resulting $n \times 2$ table the Pearson chi-square statistic X^2 and the likelihood ratio statistic (deviance) D can be computed by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)} \quad \text{and} \quad D = -2 \left(\sum_{i=1}^{n} y_i \ln \left(\frac{y_i}{\hat{\pi}_i} \right) + (1 - y_i) \ln \left(\frac{1 - y_i}{1 - \hat{\pi}_i} \right) \right).$$

We cannot use the large sample chi-square approximation to the distribution of X^2 and D since the number of covariate patterns increases with the sample size and hence the estimated expected frequencies are small. However a large sample normal approximation for the distribution of X^2 has been derived by Osius and Rojek (1992).

Hosmer and Lemeshow (2000, p. 153) describe how this approximation can be implemented in statistical software packages. The procedure simplifies if external validation is used. First the parameters are estimated from the training sample; these parameter estimates are used to calculate predicted probabilities for the observations in the validation sample. Then the Pearson chi-square statistic X^2 is calculated for the validation sample and standardized by

$$Z = \frac{X^2 - n_v}{\sigma_v} \ ,$$

where n_v is the total number of observations in the validation sample and $\sigma_v^2 = \sum_{i=1}^{n_v} \frac{1}{\hat{\pi}_i(1-\hat{\pi}_i)} - 4n_v$ (Hosmer and Lemeshow (2000, p. 187)). Osius and Rojek (1992) have shown that Z has a large sample standard normal distribution; they recommend to use a two tailed p-value. Their results do not apply to the likelihood ratio statistic D.

3.5.2 Stukel's Test

One of the basic assumptions of the logistic regression model is that the logit transformation (3.1) is the correct function linking the linear predictor with the conditional mean $\pi(\mathbf{x}_i)$. A test of this assumption can be derived by considering a generalization of the logistic regression model proposed by Stukel (1988). Two shape parameters α_1 and α_2 are introduced in order to allow for asymmetric probability curves with heavier or lighter

tails than the logistic distribution. The usual logistic regression model corresponds to the case $\alpha_1 = \alpha_2 = 0$.

A test for the hypothesis that the logit link is the correct link against the alternative of link function that results in heavier and/or lighter tails can be obtained from these ideas. First, two new variables based on the estimated logit $\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$ are introduced:

$$z_1 = 0.5 \times \left(\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)\right)^2 \times I(\hat{\pi}_i \geq 0.5) \text{ and } z_2 = -0.5 \times \left(\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)\right)^2 \times I(\hat{\pi}_i < 0.5).$$

Then the score test for addition of z_1 and z_2 to the model is performed, which has a large sample chi-square distribution with two degrees of freedom. This test is equivalent to the Score test that $\alpha_1 = \alpha_2 = 0$ in a generalized logistic regression model. A simulation study by D.W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow (1997) indicated that this test has moderate power to detect an asymmetric link function.

3.5.3 Hosmer-Lemeshow Test

Hosmer and Lemeshow (2000, p. 147) have proposed to collapse the $J \times 2$ contingency table introduced in section 3.5.1 to a $g \times 2$ table by grouping the observations according to the estimated probabilities $\hat{\pi}_i$. The grouping can be based either on fixed values or on percentiles of the estimated probabilities. For the first method, we choose cutpoints k/g, $k = 1, \ldots, g-1$, and the groups contain all subjects with estimated probabilities between adjacent cutpoints. For the second method, the first group contains the n/g subjects having the smallest estimated probabilities and so on.

The observed frequencies for the two columns of the table are the number of positive responses ($y_i = 1$) and the number negative responses ($y_i = 0$) in each group. The estimated expected frequencies are found by summing the estimated probabilities of

success ($\hat{\pi}_i$) and failure ($1 - \hat{\pi}_i$) for all g groups. Then the Pearson chi-square statistic is calculated from the $g \times 2$ table of observed and estimated expected frequencies. The Hosmer-Lemeshow statistic is defined by

$$\hat{C} := \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

where n'_k denotes the number of observations, o_k the number of positive responses and $\bar{\pi}_k$ the average estimated probability in group k.

If the fitted logistic regression model is the correct model, then the distribution of $\hat{C}$ for both types of grouping is well approximated by the chi-square distribution with $g-2$ degrees of freedom. The approximation depends on the assumption that the estimated expected frequencies are large (≥ 5). If any of the estimated expected frequencies is smaller than 5, then adjacent rows should be combined until the condition is satisfied; the number of degrees of freedom has to be reduced accordingly.

The grouping method based on percentiles results in better adherence to the χ^2_{g-2} distribution and is therefore preferred. A common choice is $g = 10$, in which case the groups are sometimes referred to as the deciles of risk. For example, the first group represents the 10% of the contracts with the smallest estimated cancellation probability and so on.

If the Hosmer-Lemeshow test statistics are computed for the validation sample, then each term in the definition of $\hat{C}$ has a large sample chi-square distribution with one degree of freedom. For g groups the distribution of $\hat{C}$ for the validation sample is approximately χ^2_g. In addition to computing a p-value for the overall statistic $\hat{C}$, the individual terms can be examined separately.

The 10×2 table itself can serve as a useful overall summary of lack of fit that is easily understood by suject matter scientists. This explains why the Hosmer-Lemeshow tests have become very popular tools and are implemented in several computer packages. However, further research has shown that there are disadvantages in the use of fixed groups based on estimated probabilities. Some issues are illustrated in a paper by Hosmer et al. (1997): First of all, the value of the test statistic depends on the choice of the cutpoints. There are examples where one set of fixed groups shows that the model fits while the tests rejects fit using a different set of fixed groups. In addition, these tests may have low power for detecting certain types of lack of fit. The grouping based on estimated probabilities ("y-space") results in groups that may contain subjects with widely different values of the covariates ("x-space").

There has been considerable work on the development of alternative methods. Hosmer et al. (1997) compared some of these alternative approaches via simulation studies and recommend to use a combination of tests: the normal approximation to the Pearson chi-square statistic for power against overall non-linearity on the logit and Stukel's test for power against a non-logit link. Despite its drawbacks, they also use the Hosmer-Lemeshow deciles of risk test for confirmatory evidence. I will follow their recommendations.

3.5.4 Classification Tables and Area under the ROC Curve

When assessing the goodness of fit of a model we should keep in mind the main purpose of the model, which in this study is to predict the outcome accurately. The response is either cancelled or not cancelled. From the fitted model a predicted

probability of cancellation can be calculated for each contract. If the predicted probability exceeds some cutpoint $c \in [0, 1]$, then the contract is predicted to be cancelled; otherwise it is predicted as not cancelled. A 2×2 classification table is obtained by crossclassifying observed and predicted outcomes. The accuracy of the classification rule can be measured by the overall misclassification rate, and by sensitivity (proportion of cancelled contracts that are classified as cancelled) and specificity (proportion of not cancelled contracts that are classified as not cancelled).

If the same observations used to fit the model are also used to estimate the misclassification rate, the resulting estimate is biased. One solution is to use external validation, that is, to set up a classification table for the validation sample based on the parameter estimates from the training sample. If the sample size is too small such that external validation cannot be used, the bias can be reduced by the use of crossvalidation. Leave-one-out crossvalidation, for example, removes one observation to be classified from the data, reestimates the parameters and classifies the observation based on the new parameter estimates. This process is repeated for every observation in the dataset. For details on crossvalidation see Hastie et al. (2000, p. 214).

Misclassification rate, sensitivity and specificity depend on a single cutpoint c. We can get a more complete description of the ability of the model to classify observations correctly by the area under the Receiver Operating Characteristic (ROC) curve. We obtain this curve by calculating sensitivity and specificity for a range of possible cutpoints and plotting sensitivity versus (1 - specificity).

The area under this curve provides a measure of discrimination with the following interpretation (Hosmer and Lemeshow (2000, p. 162)): suppose our sample consists of

n_1 cancelled and n_0 not cancelled contracts. Then consider the $n_1 \times n_0$ pairs obtained by pairing each cancelled contract with each not cancelled contract. We calculate the number of pairs for which the cancelled contract has a higher predicted cancellation probability than the not cancelled contract. When the probability is the same we just add 1/2. The area under the ROC curve is equal to the resulting number divided by the number of pairs. Hosmer and Lemeshow also note that the count we get above is equal to the Mann-Whitney U statistic for these data $(y_i, \hat{\pi}_i)$. They give the following guidelines for interpretation:

- $ROC = 0.5$: no discrimination (we might as well flip a coin)
- $0.7 \leq ROC < 0.8$: acceptable discrimination
- $0.8 \leq ROC < 0.9$: excellent discrimination
- $ROC \geq 0.9$: outstanding discrimination

The calculations can also be used to find an optimal cutpoint for the purpose of classification. If both types of error are supposed to be weighted equally, we may select the cutpoint such that sensitivity equals specificity.

Some caution is necessary when using sensitivity, specificity, misclassification rate and area under the ROC curve as measures of goodness of fit. They all measure the ability of the model to discriminate between the two outcomes. It is possible that a model discriminates well between the two outcomes but is not well calibrated. That is, the probabilities do not reflect the true outcome experience. We should always consider summary measures like the Pearson chi-square statistic to investigate the calibration of the model, especially if one of the goals is to interpret the coefficients of the model.

3.5.5 Other Measures

Considerable effort has been made to establish summary measures of goodness of fit for logistic regression models that are comparable to the R^2 measures in linear regression. Hosmer and Lemeshow (2000, p. 164) describe several of these measures and also refer to a comprehensive study of these measures of explained variation. The problem is that all these measures give low values when compared to R^2 measures usually encountered in linear regression. Therefore they are not very useful in making global statements about the goodness of fit of a model. However, they may be helpful in the evaluation of competing models.

Both Cox and Snell (1989) and Hosmer and Lemeshow (2000) recommend to use regression diagnostics in addition to the overall measures described above. Residuals and other statistics based on individual observations can help to detect anomalous or influential observations or unexpected patterns. Since the sample size for this study is large (20,000 records), the application of these measures is limited. However, it may sometimes be useful to inspect the individual components of overall measures like the Pearson chi-square statistic. For a detailed discussion of regression diagnostics see Cox and Snell (1989, p. 69) and Hosmer and Lemeshow (2000, p. 167).

3.6 Interpretation of the Model

After assesssing the goodness fit, we come to the last step in the data analysis procedure, the interpretation of the selected model. The question is how the model helps to understand the dependence of the outcome "cancellation" on the covariates $\mathbf{x}_i$.

Three cases can be distinguished: binary covariates, polychotomous covariates, and continuous covariates. A more detailed discussion that also involves interactions can be found in Hosmer and Lemeshow (2000, chap. 3).

First consider a single binary covariate x with possible values 0 and 1 and a logistic regression model of the form $\text{logit}(\pi(x)) = \beta_0 + \beta_1 x$. The odds of cancellation among contracts with $x = 1$ is defined as $\frac{\pi(1)}{1-\pi(1)}$. Similarly, the odds of cancellation among contracts with $x = 0$ is defined as $\frac{\pi(0)}{1-\pi(0)}$. The odds ratio is defined to be the ratio of the odds of cancellation for contracts with $x = 1$ to the odds of cancellation for contracts with $x = 0$. It is given by the equation

$$\text{OR} := \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}$$

and can be interpreted in the following way: if $OR = r$, then the odds of cancellation for contracts with $x = 1$ are r times the odds of cancellation for contracts with $x = 0$ (see Agresti (1990, p. 14)).

The odds ratio is a widely used measure of association for contingency tables (especially 2×2 tables) because its value does not change if rows or columns in the 2×2 table are multiplied by a nonzero constant. That means that the odds ratio can be used for interpretation even if the sample is unbalanced in some rows/columns. For example, in this study we have an equal number of cancelled and not cancelled contracts in the sample which does not reflect the actual proportion of cancelled contracts. However, the sample odds ratio we obtain by crossclassifying the outcome and a binary covariate estimates the same quantity as the sample odds ratio obtained from a sampling design

that samples cancelled and not cancelled contracts according to their actual proportion in the portfolio (see Agresti (1990), p. 16).

Another advantage of the odds ratio is that it approximates the relative risk. The relative risk is defined by

$$\text{Relative Risk} := \frac{\pi(1)}{\pi(0)}$$

and can be interpreted in the following way: if the relative risk is equal to r, then contracts with $x = 1$ are r times as likely to be cancelled than contracts with $x = 0$ (see Agresti (1990), p.14)). This is a very simple and easily understandable description of the dependence of the outcome on x. It follows from these definitions that the odds ratio approximates the relative risk if $\frac{1-\pi(0)}{1-\pi(1)} \approx 1$. This the case if both $\pi(0)$ and $\pi(1)$ are close to zero, that is, if the probability of cancellation is small for both groups formed by the binary covariate (see Agresti (1990), p. 17)).

For the logistic regression model above the relationship between the odds ratio OR and the regression coefficient β_1 is $OR = e^{\beta_1}$. This simple relationship is one of the reasons why logistic regression are considered a very useful data analysis tool. The odds ratio is estimated by plugging in the parameter estimate $\hat{\beta}_1$ in the formula above. A confidence interval for the odds ratio can be obtained by first calculating the endpoints of a confidence interval for the coefficient β_1 (using the asymptotic normality of the maximum likelihood estimate), and then exponentiating these endpoints. The resulting confidence intervals are not symmetric about the point estimate $e^{\hat{\beta}_1}$ (see Hosmer and Lemeshow (2000, p. 52)).

Now suppose we have an independent variable with k levels. The interpretation of the coefficients depends on the coding of the design variables used in the logistic regression model. If reference cell coding is used, then the coefficients are easily interpretable in a manner similar to the binary case (see Hosmer and Lemeshow (2000, p. 57)).

Reference cell coding means that we choose a reference group among the k groups formed by the covariate. The $k-1$ design variables are all set to zero for this group, and for all other groups one of the design variables is set equal to one and all others equal to zero. Then the odds ratio for every group relative to the reference group can be estimated by exponentiating the corresponding coefficient in the logistic regression model. We can obtain confidence intervals in the same way as described above. The choice of the reference group depends on the specific meaning of the covariate. An alternative would be effect coding, which results in estimates for the odds for one group relative to the geometric mean of the odds for all groups. For details see Hosmer and Lemeshow (2000, p. 59).

The interpretation of the coefficient of a continuous covariate depends on the particular units of the variable. Exponentiation of the parameter estimate yields an odds ratio for a one-unit change in the continuous covariate. Very often a one-unit change is not very meaningful, but a change in the continuous covariate of c units makes sense in the context of the problem. Then an estimate for the odds ratio for a change of c units can be obtained by computing $e^{c\hat{\beta}_1}$; confidence intervals are computed similarly.

Sometimes a logistic regression model is of the form $\text{logit}(\pi(x)) = \beta_0 + \beta_1 d + \beta_2 x$, where $d = 0$ if $x > 0$ and $d = 1$ if $x = 0$. This is reasonable if the value 0 has a special meaning that distinguishes it from positive values for x. Under this parameterization,

two types of odds ratios can be estimated. For positive values of x, the odds ratio corresponding to a change of c units is estimated by $e^{\hat{\beta}_2 c}$ as discussed above. Additionally, we can estimate the odds ratio comparing observations with a value of $x_2 > 0$ versus observations with $x_1 = 0$ by $e^{\hat{\beta}_2 x_2 - \hat{\beta}_1}$. In order to obtain a confidence interval for this estimated odds ratio, we first estimate the standard error of the linear predictor $\hat{\beta}_2 x_2 - \hat{\beta}_1$ by $[x_2^2 \widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_1) - 2x_2 \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_1)]^{(1/2)}$, then compute a confidence interval for $\hat{\beta}_2 x_2 - \hat{\beta}_1$ using the normal approximation, and finally exponentiate the endpoints.

CHAPTER 4

CLASSIFICATION TREES

For the discussion of classification trees in this chapter I will closely follow the methodology introduced by Breiman, Friedman, Olshen, and Stone (1984).

4.1 Basic Idea

The setup for a classification tree is as follows: For every subject in a study we have measured a vector of explanatory variables (covariates) $\mathbf{x_i}$, together with a classification y_i into one of K classes. The set of all possible vectors $\mathbf{x_i}$ is called the covariate space $\mathcal{X}$, and the set of possible classes is denoted by $\mathcal{C}$.

A (binary) tree structured classifier is then constructed by repeated (binary) splits of subsets of the covariate space $\mathcal{X}$, beginning with the complete covariate space. Some subsets are not split, and are called terminal subsets; they form a partition of the covariate space. For each terminal subset a class label is assigned, where different terminal subsets can have the same class label. The partition is gotten by putting together all terminal subsets corresponding to the same class. Thus, classification trees are a hierarchical way of describing a partition of $\mathcal{X}$.

The terminology of trees is graphic. A node t corresponds to a subset of $\mathcal{X}$; the root node t_1 corresponds to the complete covariate space. Terminal subsets become terminal nodes or leaves and non-terminal subsets are called non-terminal nodes.

Each non-terminal node contains a question on which the split is based. Any possible observation in the covariate space can be passed down the tree, with decisions being made at each node until a terminal node (leaf) is reached. The observation is then classified according to the corresponding class label.

The construction of such a tree is based on past experience, which is summarized by a learning sample. The learning sample consists of the covariate vectors $\mathbf{x_i}$ for N cases observed in the past and their actual classification y_i. This sample is used to construct a classification tree, where the construction is centered on three major elements (see Breiman et al. (1984, p. 22)):

1. The selection of the splits.
2. The decision when to declare a node terminal or to continue splitting it.
3. The assignment of each leaf to a class.

I will discuss the first two aspects in more detail in the following sections; this will also include a solution to the third question.

The discussion will be focused on the methodology introduced by Breiman et al. (1984) and implemented in the CART (Classification And Regression Trees) software package. I will also refer to more recent developments discussed in Ripley (1996). Other approaches that should be mentioned are CHAID (see Kass (1980)) and C4.5/C5.0 (see Quinlan (1993)).

4.2 Selection of Splits

Before discussing how to actually select the "best" split at a node t, we first have to define the set of possible splits. Usually only binary splits are considered; each node

is split into just two groups at each stage. Sometimes it may be useful to allow for splits into more than two groups (multiway splits). However as Hastie et al. (2001, p. 273) point out, this is not a good general strategy: First, the data are split too quickly using multiway splits; insufficient data may be left at the next level. Second, any multiway split can be achieved by a series of binary splits.

Every binary split can be expressed as a question; if the answer is yes, then an observation is sent to the left branch otherwise to the right branch. Breiman et al. (1984, p. 29) define the following "standard set of questions":

1. Consider only splits on a single variable x_m, where x_m can be measured on any type of measurement scale.
2. Possible questions for a categorical variable with levels $b_1, \ldots, b_L$ are of the form $\{\text{Is } x_m \in S?\}$, where S can be any subset of $b_1, \ldots, b_L$.
3. Possible questions for a continuous variable are of the form $\{\text{Is } x_m \leq c?\}$, where c ranges over $(-\infty, \infty)$.

The set of possible splits corresponding to these questions is denoted by $\mathcal{S}$ and it is finite: For a continuous variable there are as many possible splits as there are different values and for a categorical variable with L levels, there are $2^{L-1} - 1$ different splits.

This standard set can be extended to allow for splits based on linear combinations of variables (see Breiman et al. (1984, p. 38 and p. 132)). In order to discover and use linear structure in the data, splits of the form $\sum_m a_m x_m \leq c$, where $\sum_m a_m^2 = 1$, are included in $\mathcal{S}$. An effective search algorithm to find the best split is presented. The disadvantage is that there is a loss in interpretability and that splits are no longer invariant under monotone transformations of variables x_m. Breiman et al. (1984, p. 136) also discuss the use of splits based on boolean combinations of variables.

The next step is to describe how the "best" split at a node t can be found. For a better understanding it is helpful to assume that we have a probability distribution over $\mathcal{X} \times \mathcal{C}$. By partitioning, we can obtain $p(k,t)$, the probability that an observation reaches the node t and is in class k from the distribution over $\mathcal{X} \times \mathcal{C}$. Then $p(t) = \sum_{k=1}^{K} p(k,t)$ is the (marginal) probability that an observation falls in node t and $p(k|t) = \frac{p(k,t)}{p(t)}$ is the (conditional) probability that an observation is in class k given that it falls in node t.

All these probabilities are estimated by the respective quantities in the learning sample. For example $\hat{p}(k|t) = \frac{N(k,t)}{N(t)}$, where $N(t)$ is the number of observations in the learning sample that fall in node t and $N(k,t)$ is the number of class k observations among all observations in node t. These estimated proportions are used to assign a class label to each node by the plurality rule: assign class label $k^*(t)$ for node t if

$$\hat{p}(k^*(t)|t) = max_{k=1,\ldots,K}\hat{p}(k|t).$$

Generalizations of this rule are discussed in section 4.4.

Breiman et al. (1984, p. 23) describe the basic idea of split selection: look for a split of a subset so that the data in the descendant subsets are "purer" than the data in the parent subset. Reasonable measures of impurity should be zero if $p(k|t)$ is concentrated on one class and maximal if it is uniform on the K classes. Commonly used measures of impurity $i(t)$ of a node t are misclassification rate, Gini index and entropy (or deviance):

- Misclassification rate:

$$i(t) = \frac{1}{N(t)} \sum_{(x_i,y_i)\in t} I(y_i \neq k^*(t)) = 1 - \hat{p}(k^*(t)|t).$$

- Gini index:

$$i(t) = \sum_{k \neq k'} \hat{p}(k|t)\hat{p}(k'|t) = \sum_{k=1}^{K} \hat{p}(k|t)(1 - \hat{p}(k|t)) = 1 - \sum_{k=1}^{K} (\hat{p}(k|t))^2.$$

- Entropy (deviance):

$$i(t) = -\sum_{k=1}^{K} \hat{p}(k|t) \ln (\hat{p}(k|t)) \quad (\text{where } 0 \ln 0 := 0).$$

For a two-class problem ($C = \{0, 1\}$) with p being the proportion of class 1 observations in node t the three measures become $i(t) = 1 - \max(p, 1-p)$, $i(t) = 2p(1-p)$ and $i(t) = -p \ln(p) - (1-p) \ln(1-p)$. Additionally, if either Gini index or entropy are used in a two-class problem, then the number of possible splits for a categorical variable can be reduced to $L - 1$ (for a nice proof see Ripley (1996, p. 218)). I will discuss interpretations of these measures later in this section.

First suppose we have a candidate split s at node t which divides into nodes t_L and t_R and sends the proportion p_L to t_L and p_R to t_R. The goodness of this split is defined to be the decrease in average impurity:

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R).$$

The best split for this node is the one that maximizes the goodness of split measure. The algorithm searches through all possible splits in $\mathcal{S}$ to find the best split. Then node t is split and the nodes t_L and t_R are considered for further splitting.

For illustration, Breiman et al. (1984, p. 32) also define the overall tree impurity for a tree T with set of leaves $\tilde{T}$ by:

$$I(T) = \sum_{t \in \tilde{T}} i(t)p(t).$$

They show that maximizing the decrease in node impurity is equivalent to minimizing the overall tree impurity. The splitting procedure can therefore be seen as a repeated attempt to minimize overall tree impurity.

If the sample size at a particular node t is too large (larger than a fixed maximum), computational efficiency can be increased by using subsampling for the split search. A subsample from each class represented at node t is used to determine the split; but the entire dataset in t is sent down the split. Such a procedure usually only affects upper nodes in a tree; the sample size down the tree is not decreased (for details see Breiman et al. (1984, p. 163)). It should only be applied when the computer resources are limited.

At this point I will take a closer look at the impurity measures. The Gini index can be interpreted in two ways. First suppose that instead of using the plurality rule for assigning a class label, we choose the class label randomly from the class distribution $\hat{p}(k|t)$ at node t, that is, we classify into class k with probability $\hat{p}(k|t)$. The estimated probability that an observation is actually in class $k' \neq k$ is $\hat{p}(k'|t)$ and the estimated probability of misclassification is $\sum_{k \neq k'} \hat{p}(k|t)\hat{p}(k'|t)$, the Gini index.

A second interpretation in terms of variances is gotten by the following considerations. Assign all class k objects a value of 1 and all other objects a value of zero. The sample variance of these values is $\hat{p}(k|t)(1 - \hat{p}(k|t))$. Repeating this for all K classes and summing the variances again yields the Gini index (Hastie et al. (2001, p. 271)).

The deviance measure of impurity is based on a slightly different approach (see Venables and Ripley (1997, p. 417)). We view the tree as a probability model for the learning sample. Since each case in the learning sample is assigned to a leaf, we have a random sample of size $N(t)$ from the multinomial distribution specified by

$p(k|t), k = 1, \ldots, K$, at each leaf t. If we now condition on the vectors $\mathbf{x_i}$ in the learning sample, we know $N(t)$ and $N(k,t)$ for all nodes, in particular for the leaves. The conditional likelihood is proportional to

$$\prod_{leaves\ t} \prod_{classes\ k} p(k|t)^{N(k,t)}.$$

The deviance is then defined to be the conditional log-likelihood multiplied by -2:

$$D(T) = -2 \sum_{leaves\ t} \sum_{classes\ k} N(k,t) \ln(p(k|t)).$$

The quantities $p(k|t)$ are estimated by the maximum likelihood estimates $\frac{N(k,t)}{N(t)}$. One splitting strategy is to choose the split which maximizes the decrease in deviance. If we now consider the overall tree impurity $I(T)$ for the entropy measure, it can be shown (see Ripley (1996, p. 219)) that $I(T) = D(T)/2N$. Hence the splitting strategies based on deviances and on the entropy measure are identical.

In general, Gini index and entropy are preferred to the misclassification rate as impurity measures used for growing a tree. Breiman et al. (1984, p. 95) discuss reasons why the misclassification rate should not be used. First, the decrease in impurity may be zero for all possible splits in $\mathcal{S}$ under rather weak conditions. Second, even though minimizing the misclassification rate is the overall goal, it is not advisable to use it as a criterion for the split search, since we only do a one-step optimization.

Breiman et al. (1984) seem to prefer the Gini index, whereas the entropy is used frequently in the machine learning literature. For two-class problems, the resulting trees should be very similar (see, for example, the plot of the impurity measures in Hastie et al. (2001, p.271)); the pruning criterion seems to be much more important (see section 4.3).

For multi-class problems, however, different splitting criteria may lead to substantially different results. Therefore it is good practice to experiment with different rules. One alternative splitting criterion (twoing) for the multi-class problem is defined in Breiman et al. (1984, p. 104).

4.3 Pruning

In the previous section I discussed how to grow a classification tree by repeated binary splits on single variables. The next obvious question is when to stop the growth of a tree.

In statistics, we usually have to deal with "noisy" classification problems; the class distributions overlap and there is no exact partition of the covariate space $\mathcal{X}$. If there was one, we could simply grow the tree until every observation in the learning sample is classified correctly. Doing so in a noisy problem would over-fit the observations; the tree adapts too well to the observations in the learning sample and has a higher true misclassification rate than the smaller, right-sized tree. On the other hand, if the growth of the tree is stopped too early, we do not use all the classification information available in the learning sample and also get a higher true misclassification rate (Breiman et al. (1984, p. 60)).

Early approaches tried to find appropriate stopping rules, that is, criteria for declaring a node terminal. A simple example is to set a threshold β and to decide not to split a node if the maximum decrease in impurity is less than β. The difficulty here is how to choose β. If β is set too low, then the resulting tree is too large. Increasing β

results in another problem: There may be nodes t for which $\Delta i(s,t)$ is small for all possible splits; but for the descendants t_L and t_R, there are splits with large decreases in impurity. If t is declared terminal, the information in the good splits on t_L and/or t_R is lost (see Breiman et al. (1984, p. 61)). An example for a stopping criterion based on significance tests is the CHAID algorithm described in Kass (1980). It also did not lead to satisfactory results.

Finally these approaches were discarded in favor of a completely different way of looking at the problem that can be described as a three-step procedure: First grow a tree that is much too large. Then prune the tree upward in a reasonable way until you end up at the root node. Finally use accurate estimates of the true misclassification rate to select the right-sized tree from among the pruned subtrees (Breiman et al. (1984), p. 37).

Breiman et al. (1984, p. 62) discuss this procedure in great detail. They recommend to grow a very large tree T_0 first. Nodes are split until either

- the node is pure (all cases are in one class, impurity measure is zero) or
- the node contains only identical measurement vectors (splitting would result in one empty descendant node) or
- the node is small, that is, $N(t) \leq N_{min}$. Typical choices for N_{min} are 1 or 5, depending on the problem and the available computer resources.

Then we start with T_0 and selectively prune it upward. This results in a sequence of smaller and smaller subtrees that eventually collapses to the tree consisting only of the root node. We would like each subtree to be the "best" in its size range. Breiman et al. (1984, p. 284) showed to the number of subtrees that have the same root as T_0 is approximately $\lfloor 1.5028368^l \rfloor$, where l is the number of leaves of T_0. Hence there is large

number of distinct ways of pruning upward to the root node. I will discuss two algorithms that can be used to get such a sequence of subtrees.

Probably the simplest pruning process works as follows. Suppose T_0 has L leaves. Then for each value of H, $1 \leq H < L$, consider all subtrees of T_0 having exactly $L-H$ leaves. Among those, choose the subtree T_H that has the smallest misclassification rate $R(T)$ on the learning sample. This procedure has one drawback: the sequence of subtrees is not necessarily nested, that is, T_{H+1} is not necessarily a subtree of T_H. Nodes that were just cut off may reappear as we go through the sequence. However the process is intuitively appealing, gives the best subtree for every possible number of leaves, and can be implemented effectively (see Breiman et al. (1984, p. 65)).

The second pruning method I want to discuss is cost-complexity pruning which was first introduced by Breiman et al. (1984) and represents the established methodology. They define the cost-complexity measure of a tree T by

$$R_\alpha(T) = R(T) + \alpha * \text{size}(T),$$

where $R(T)$ is the misclassification rate on the learning sample, size(T) is the number of leaves and $\alpha \geq 0$ is the complexity parameter. According to Ripley (1996, p. 221), another possible measure $R(T)$ could be the deviance.

Then $T(\alpha)$ is defined to be the smallest minimizing subtree for given complexity parameter α, that is, $T(\alpha)$ is the smallest subtree such that

$$R_\alpha(T(\alpha)) = \min\{R_\alpha(T) : T \text{ is any subtree of } T_0\}.$$

Breiman et al. (1984) showed that for every $\alpha \geq 0$ there is such a subtree $T(\alpha)$ and that it is unique. They also noted that there is only a finite number of subtrees of T_0, but α

runs through a continuum of values. This means that if $T(\alpha)$ is the smallest minimizing subtree for some α, then it continues to be optimal as α increases until a jump point α' is reached and a new tree $T(\alpha')$ becomes minimizing. If we continue to increase α, then we finally end up at the tree consisting only of the root node.

The resulting sequence of subtrees is finite and nested. That is, each subtree can be gotten from the previous tree by pruning upward. Furthermore an algorithm, based on the so-called weakest link cutting, was developed for the effective computation of this sequence. For details see Breiman et al. (1984, p. 68) and Ripley (1996, p. 222). In the beginning of the process, the algorithm tends to prune off larger subbranches with many terminal nodes; as the algorithm produces smaller trees, the number of leaves cut off at a time tends to decrease.

These two pruning methods are closely related. If $T(\alpha)$ is the smallest minimizing subtree with respect to the cost-complexity measure (based on the misclassification rate) and has n leaves, then it is also the subtree which has the smallest misclassification rate among all subtrees with n leaves. That is, the sequence of subtrees gotten using cost-complexity pruning is a nested subsequence of the sequence of subtrees gotten by minimizing the misclassification rate for every possible number of leaves. Ripley (1996, p. 226) describes some alternatives to these pruning methods.

What remains at this point (no matter which pruning method was used) is to choose the "best" subtree within the sequence. The criterion usually used is the true misclassification rate of a tree, which of course is unknown for a noisy problem. Breiman et al. (1984, p. 72) recommend to use an "honest" estimate of the true misclassification rate and describe two ways of obtaining this estimate.

If the dataset is large, then it is split into a training and a validation sample, usually in a ratio of 2:1 or similar. The training set is used to grow the large initial tree T_0 and to prune it. The subtree with the smallest misclassification rate on the validation sample is then choosen from the sequence. Alternatively, for small datasets the use of V-fold crossvalidation is recommended.

For both external validation and crossvalidation the plot of the estimated misclassification rate versus the number of leaves has a similar shape. As the number of leaves increases, we first see a rapid initial decrease, followed by a long flat valley where the misclassification rate only fluctuates slightly, and then a gradual increase for very large trees (due to overfitting).

In order to reduce the instability and to find the simplest tree whose accuracy is "close" to the optimum, Breiman et al. (1984, p. 78) recommend to use the "one standard error rule" to select the tree. Since the n_v observations in the validation sample are independent of the observations in the training sample, we can describe the process of classifying these observations using the tree by a binomial model with n_v independent trials and some common probability of misclassification p. The proportion $R_v(T)$ of misclassified observations in the validation sample is an unbiased estimator for p. Its estimated standard error is

$$\widehat{S.E.}(R_v(T)) = \sqrt{\frac{R_v(T)(1 - R_v(T))}{n_v}}.$$

The smallest subtree in the sequence whose misclassification rate is just within one estimated standard error of the minimum is selected.

The methods discussed in the two previous sections represent the basic established methodology for classification trees. There are some features that have been added to the basic tree structure in order to make it more flexible and more powerful. One of them is the use of priors and the inclusion of misclassification costs. A second one is the use of surrogate rules for the handling of missing values and for assessing the importance of the individual covariates. They are discussed in the two remaining sections of this chapter.

4.4 Priors and Misclassification Costs

Due to the sampling design, the class proportions in the learning sample sometimes do not reflect the class proportions in the whole population from which the sample was chosen. This fact can be incorporated into the tree construction process by specifying prior probabilites $\pi(k)$ for the K classes.

In this concept, the priors are interpreted as the probabilities that a class k observation is presented to the tree. They are either supplied by the analyst or estimated by the class proportions $\frac{N(k)}{N}$ in the learning sample. The estimated probability that a case is in class k given that it falls in node t is now $\hat{p}(k|t) = N\frac{\pi(k)}{N(k)} \times \frac{N(k,t)}{N(t)}$, which reduces to $\frac{N(k,t)}{N(t)}$ for $\pi(k) = \frac{N(k)}{N}$. The same class assignment rule as in section 4.2 is used with the modified estimates $\hat{p}(k|t)$ (see Breiman et al. (1984, p. 34 and p. 112)).

Priors can be used to adjust the individual class misclassification rates in any desired direction. Equal priors tend to equalize the misclassification rates, whereas a larger prior on one class tends to decrease the misclassification rate for this class.

Using priors $\pi(k)$ is equivalent to weighting observations in class k by weights $N\frac{\pi(k)}{N(k)}$ (see Ripley (1996, p. 220)).

So far the assumption has been made that the cost or loss in misclassifying class k objects as class k' objects is the same for all $k \neq k'$. In some classification problems, the consequences of misclassifying observations are more serious in some classes than in others. If this is the case, we can include a set of misclassification costs $C(k'|k)$, where $C(k'|k)$ is the cost of misclassifying a class k object as a class k' object. We assume that $C(k'|k) \geq 0$ for all $k \neq k'$ and $C(k'|k) = 0$ for $k = k'$.

These costs can then be included in the class assignment rule. If an observation of unknown class is selected randomly, falls into node t and is classified as class k', then the estimated expected misclassification costs at that node are $\sum_{k=1}^{K} C(k'|k)\hat{p}(k|t)$. The class label $k^*(t)$ of node t is the value of k' that minimizes this sum (see Breiman et al. (1984, p. 35)).

Using the learning sample, the expected misclassification cost at node t can be estimated by $r(t) = \min\left\{\sum_{k=1}^{K} C(k'|k)\hat{p}(k|t) : \ k' \in \{1, \ldots, K\}\right\}$ and the estimated expected misclassification cost of the tree T with leaves $\tilde{T}$ is given by $R(T) = \sum_{t\in\tilde{T}} r(t)p(t)$. $R(T)$ is then used in the pruning process; this is why $R_\alpha(T)$ is called cost-complexity measure. In the case of unit cost ($C(k'|k) = 1$ for all $k \neq k'$), $R(T)$ simplifies to the misclassification rate on the learning sample.

Misclassification costs may also be included in the splitting rule (see discussion in Breiman et al. (1984, p. 113)). A direct inclusion in the definition of the Gini index is completely ineffective in two-class problems because it symmetrizes costs. Instead of including the costs in the definition of the impurity measure, Breiman et al. recommend

to adjust the priors in order to take into account varying misclassification costs. This works perfectly for two-class problems and also for multi-class problems if, for each class, there is a constant misclassifiction cost that does not depend on how an observation is misclassified ($C(k'|k) = C(k), k \neq k'$). Ripley (1996, p. 221) discusses how misclassification costs and priors can be incorporated in the deviance measure by using weights.

4.5 Surrogate Splits

Consider a node t with optimal split s^* that sends the proportion p_L to the left branch t_L and p_L to the right branch t_R. The goal is now to predict the action of s^*. A simple rule would be to predict t_L, if $p_L = \max(p_L, p_R)$ and t_R otherwise. That is, we determine the branch to which most observations are sent using s^* and simply send all observations to that branch. Obviously the error probability of this rule is $\min(p_L, p_R)$.

We may be able to do better by applying the following procedure. For any variable x_m find the split $\tilde{s}_m$ on x_m that most accurately predicts the action of s^*. That is, $\tilde{s}_m$ is the split on x_m at node t which sends the largest proportion of observations in t to the same branch as s^* would do. $\tilde{s}_m$ is called a surrogate split on x_m for s^* at node t (Breiman et al. (1984, p. 140)). Such a rule $\tilde{s}_m$ only makes sense if its error rate is smaller than $\min(p_L, p_R)$, the error rate of the simple rule discussed above.

One application of surrogate rules is an intelligent way of handling missing values. All observations, even those with missing values, can be used in a natural way in the tree construction. If there are missing values for some of the cases at a node t, then first consider each variable x_m in turn and find the split s_m^* on x_m using all observations

without a missing value for x_m. Among those splits s_m^* for all variables x_m, choose the split s^* that results in the largest decrease in impurity.

Then for a particular observation, the split s^* at node t may not be defined because the observation has a missing value. If this is the case, consider all nonmissing variables for this observation and find the one, say x_l, for which the surrogate split has the smallest error rate. Assign the observation to a branch using $\tilde{s}_l$ (Breiman et al. (1984, p. 142)).

This algorithm works much better than most algorithms for missing values in regression models. We make the most possible use of the data to grow the tree and the worst case is that the smallest error probability of a surrogate rule is close to $\min(p_L, p_R)$. But even if an observation is sent to the wrong branch by the surrogate rule at some node, it may still be classified correctly since the splitting continues below that node.

A simple alternative to the described way of handling missing values is to drop an observation as far as it goes in the tree and use the highest proportion in that node for the classification. Sometimes the fact that a value is missing carries some information. Then (and only then) it is reasonable to treat "missing" as a seperate value for a variable and use it in the split search. Ripley (1996, p. 232) discusses these approaches in some detail.

Surrogate rules can also be used to get an idea of the relative importance of the covariates used for splitting. To measure the importance of a variable x_m, find the surrogate split $\tilde{s}_m$ on x_m for all nodes t and compute the decrease in tree impurity $\Delta I(\tilde{s}_m, t)$

if $\tilde{s}_m$ is used at node t. Breiman et al. (1984, p. 147) define a measure of importance by

$$M(x_m) = \sum_{t \in T} \Delta I(\tilde{s}_m, t).$$

The concept behind this definition is the following: Suppose that the best split at a node t is on x_{m_1} and that x_{m_2} can generate a split $s^*_{m_2}$ similar to $s^*_{m_1}$ in terms of decrease in impurity. Then at node t, the decrease in impurity of the surrogate split on x_{m_2}, $\Delta I(\tilde{s}_{m_2}, t)$, will be nearly as large as $\Delta I(s^*_{m_1}, t)$. Using $\Delta I(s^*_{m_2}, t)$ in the summation for the importance measure can lead to misleading results (see discussion in Breiman et al. (1984)). Since only relative differences are important, this measure is normalized on a 0 - 100 scale.

Alternatively, the decrease in impurity may be discounted by an agreement measure which is one if $\tilde{s}_m$ is actually the primary split at node t. If $\tilde{s}_m$ is a surrogate split at node t, then the agreement is equal to the proportion of observations that is sent to the same branch by $\tilde{s}_m$ and the primary split. The measure is set to zero if the surrogate split $\tilde{s}_m$ has an error rate greater than $\min(p_L, p_R)$. This corresponds to the implementation in SAS and is also available in the CART software.

CHAPTER 5

RESULTS

This chapter contains details and results of the model building process for both logistic regression models and classification trees.

5.1 Logistic Regression Models

5.1.1 Variable Screening

The first step in the model building process should be a careful univariate analysis of each variable that seems to be a reasonable predictor considering the context of the study. The goal is to get some understanding about the dependence of the outcome on each one of the possible covariates and about the best way the covariates should be entered into the logistic regression model.

In this step we can also check for coding errors and missing values in the dataset. Before I started the analysis, the dataset had been cleaned up: values that did not make sense (for example, values that were out of the range of a variable or character values instead of numerical values) were deleted.

If a variable has missing values and it is used for the logistic regression model, then the sample size needs be reduced accordingly. Missing values do not represent a serious problem for most of the possible covariates. For all covariates except KUNALTER2 (age of the policy holder) and KAUFDAT1 (year of purchase of the vehicle), the

number of missing values is very small. For KUNALTER2, 1,935 (9.68%) values are missing. However, the age of the policy holder seems to be an important covariate. For KAUFDAT1, 9,331 values (46.66%) are missing. Using this variable directly would lead to an undue decrease in sample size. Further investigation has shown that even if "missing" is defined as a seperate category, this variable is not a useful predictor. If all observations with missing values for any of the important covariates are ignored, the remaining dataset has 18,014 records and is still balanced between cancelled and not cancelled contracts (49% cancelled, 51% not cancelled).

I have screened all 27 reasonable covariates discussed in chapter 2 using the SAS computer package (procedures FREQ, UNIVARIATE, GPLOT, LOGISTIC and CORR). For details on the use of the SAS system for categorical data analysis see Stokes, Davis, and Koch (2000) or Cody and Smith (1997). The methods used and some results are discussed below; for more details see appendix B.

5.1.1.1 Categorical covariates. For categorical explanatory variables (nominal or ordinal scale) with k levels, we first set up a $k \times 2$ contingency table. Particular attention should be paid to contingency tables with zero cells or cells with small counts because they may result in estimation problems (Hosmer and Lemeshow (2000), p. 135). One solution to these numerical problems is to combine categories in some sensible fashion, taking into account background information about the variable.

In the second step, $k-1$ dummy variables representing the levels of the categorical covariate are defined using either reference cell or effect coding and a univariate logistic regression model is fitted. The parameter estimates can be used to estimate odds ratios

which can give hints about the differences between the categories. If significance tests indicate that a variable is not significant, then this variable is not very likely to be a good predictor. Variables that are measured on ordinal scale may be modeled as if they were continuous by using scores for the levels of the variable.

For the categorical variables PLZ (zip code), DECKAR (type of coverage motor TPL), TGR (tariff group motor TPL), KZWF (mileage per year) and FANZ1 (number of drivers) the counts for some (or all) categories are small and would result in numerical problems. Therefore I decided to combine categories (taking into account background information) and defined new variables PLZX, DECKAR2, TGR2, KZWF2 and FANZ4. For details and the definition of the new variables see appendix B.

Background information and/or frequency counts also suggest to combine categories for the variables KZKHT, KZWF2 and FANZ4, which turned out to be useful in the process of model selection.

Finally, the following categorical variables may be used as predictors: PLZX (zip code, 23 levels), REGBE (county, 9 levels), KZKHT (tariff generation motor TPL, 6 levels), DECKAR2 (type of coverage motor TPL, 3 levels), REGIO (regional classification, 10 levels), TGR2 (tariff group, 4 levels), KZWF2 (mileage restriction, 4 levels), FANZ4 (restriction on number of drivers, 5 levels). Several binary variables also seem to be useful predictors for the logistic regression model, namely AUSKL1 (foreigner?), KZEHB1 (home owner discount?), KZFV (vehicle insurance?) , SORAB1 (special discount?) and VERBU1 (employee tariff?).

5.1.1.2 Continuous covariates. For continuous variables (interval or ratio scale), first descriptive statistics like mean, median and quantiles are computed; they give useful information about the range of these variables.

Then the values are combined into groups in some sensible way, preferably using the available background information. If the number of possible values is small, then the values themselves can be used as groups. After grouping the continuous variable, both the empirical proportion of cancelled contracts and the empirical logit are calculated for each group. A plot of the logits against the group values can be used to check whether the variable can be treated as linear on the logit scale. If not, we may consider polynomial terms or other transformations of the continuous covariate (for example log or square-root). As for categorical variables, the univariate logistic regression model is fitted and both parameter estimates and significance tests are examinated.

It turns out that it is reasonable to treat the following variables as linear on the logit scale: VWDAU (number of years in the portfolio), KUNALTER2 (age of the policy holder), ZART (payment option), STAERKE (vehicle power), FALTER (age of the vehicle), TYPKLH (type class motor TPL), TARKLA2 (no-claims class motor TPL), BEITSAT3 (percentage rate motor TPL), JTBH (annual premium motor TPL), TYPKL (type class vehicle insurance), TARKLAFV (no-claims class vehicle insurance), FELD68 (percentage rate vehicle insurance) and JTBF (annual premium vehicle insurance). For the variables STAERKE, JTBH and JTBF a quadratic term may also be reasonable.

Additionally to the linear term for the age of the vehicle (FALTER), it also seems useful to define a new binary variable FALTER0 that indicates whether the vehicle is new

(FALTER = 0) or not: contracts for new vehicles are more likely to be cancelled than contracts for older vehicles.

A similar approach is used for the percentage rate for the motor TPL (BEITSAT). Contracts with percentage rate above 125 (classes S, M and 0) are very likely to be cancelled, whereas there seems to a linear relationship on the logit scale for percentage rates between 30 and 125. It turned out to be useful to define a new variable BEITSAT2 that indicates whether the percentage rate is greater than 125 or not and another variable BEITSAT3, which is defined by replacing all values above 125 in BEITSAT with zero.

The variable TARKLA2 (TARKLFV) represents the no-claims classes SF 1 through SF 34 for motor TPL (vehicle insurance). For the remaining classes SF 1/2, 0, S and M in motor TPL, indicator variables were defined (TARKLA_0, TARKLA_12, TARKLA_S and TARKLA_M). For the vehicle insurance, two indicator variables for the classes SF 1/2 and 0 are used (TARKLAFV_0 and TARKLAFV_12).

5.1.1.3 Associations between predictor variables. Another important result of the variable screening is that possible values for several variables related to motor TPL insurance depend on the tariff generation (variable KZKHT), namely TYPKLH, DECKAR2, KZWF2, FANZ4 and KZEHB1. Therefore the effect of these variables on the cancellation behavior may depend on the tariff generation. This could be accounted for by using interaction terms between KZKHT and any of these four variables.

Additonally, there are two pairs of variables that contain very similar information: the percentage rate for motor TPL (variable BEITSAT) and the no-claims class for motor TPL (variable TARKLA2 with 4 indicator variables). There is a similar pair

for motor vehicle insurance (FELD68 and TARKLAFV with 2 indicator variables). TARKLA2 (TARKLAFV) represents the number of years without a claim, and the indicator variables represent claims classes (S and M) and other special classes (SF 1/2 and class 0). The values of these variables determine the percentage rate BEITSAT (FELD68), as defined in the conditions of the contract. Different no-claims classes can have the same corresponding percentage rate (for details see section 1.2). The percentage rate is then used to find the annual premium JTBH (JTBF). The decision which variables are actually entered in the logistic regression model will be made in the model selection process.

Apart from that, there are several continuous covariates that show a high level of association. Considering the Pearson correlation coefficient as a measure of linear association, it turns out that vehicle power (STAERKE), type class for motor TPL (TYPKLH) and premium for motor TPL (JTBH) are correlated, where the intensity of the association depends on the tariff generation for TPL insurance (KZKHT). Also percentage rate and premium for vehicle insurance (FELD68 and JTBF) and age of the vehicle and premium for vehicle insurance (FALTER and JTBF) seem to be correlated. These assocations suggest to consider the corresponding interaction terms in the modeling process and will limit the interpretability.

5.1.2 Model Selection and Model Assessment

As a first step in the model building process I considered the complete dataset and used stepwise procedures, namely backward elimination and forward selection, to get an idea of the relative importance of the covariates specified above. Both procedures

are based solely on statistical significance as selection criterion.

Following the recommendations in Hosmer and Lemeshow (2000, p. 92), I started with a model containing all main effects that seemed useful according to the variable screening in the previous section (a total of 37 effects). The decisions for entering or removal were based on the results of both forward selection and backward elimination; it was also useful to take into account background information about the variables.

First of all, the variables representing the percentage rates for both liability and vehicle insurance (BEITSAT and FELD68) seem more useful than the variables corresponding to the no-claims classes (TARKLA2 and TARKLAFV with indicator variables). Hence I only used BEITSAT and FELD68 in subsequent modeling steps.

Another four variables were suggested for removal from the model: AUSLK1 (foreigner?), REGBE (county), SORAB1 (special discount?) and TYPKL (type class vehicle insurance). For the binary variables AUSLK1 and SORAB1, the number of positive responses is small (<6%) and they are not considered important risk factors for cancellation. Regional classification into 23 areas is represented by the zip code (PLZX); since counties correspond to larger areas, REGBE can be removed from the model. The variable TYPKL also does not seem to give important information on the cancellation behavior, even though it is one of the key tariff factors determining the premium for vehicle insurance. The remaining 25 main effects seem to be useful (p-values < 0.05), at least in terms of statistical significance.

In the second step, I investigated whether any of the interactions mentioned in the previous section seems to be important for the prediction model. Again I used the results of backward elimination and forward selection procedures as criteria.

The result is that all six two-way interactions between STAERKE, TYPKLH, JTBH and KZKHT are significant, together with two three-way interactions. At this point it also seems reasonable to use only linear terms for STAERKE, JTBH and JTBF. Additionally, the variables FELD68 (percentage rate vehicle insurance) and DECKAR2 (type of coverage motor TPL) are not significant any more once interaction terms are entered. Considering all terms that are statistically significant and reasonable from the context, we end up with a model with 28 effects.

At this stage, I checked whether the model fits the data using external validation. The dataset was randomly divided into a training sample (2/3) and a validation sample (1/3) and p-values were computed for the goodness of fit statistics discussed in section 3.5. The corresponding SAS programs are given in appendix C. In order to confirm results I considered several (for this model, 10) of these random partitions .

Overall, both the Pearson chi-square test and Stukel's test indicate that the model fits, although the p-values fluctuate depending on the partition. For one partition, the p-value for Pearson's test is even less than 0.05. This result will be discussed in detail later.

In addition, the p-values for the Hosmer-Lemeshow tests are small (< 0.05) for some other partitions, but they seem to depend on the way the data are grouped (fixed cutpoints or percentiles). Sometimes one grouping strategy results in a p-value that suggests that the model fits, while another way of grouping leads to a p-value well below 0.05. This illustrates the disadvantages mentioned in 3.5.3 related to the use of fixed groups based on estimated probabilites.

Furthermore, I investigated the discriminatory power of the model by setting up classification tables for several cutpoints. Choosing the cutpoint such that sensitivity equals specificity resulted in misclassification rates (on the validation sample) between 15.55% and 17.23% for the 10 random partitions, where the median was 16.63%. The area under the ROC curve varied between 0.914 and 0.925 with a median of 0.920.

The conclusion is that this model provides excellent discrimination, which is the primary goal. However, it seems to be rather complex (eight interaction terms and a total of 68 parameters) and the question is whether a simpler model is capable of predicting the outcome with comparable accuracy while still fitting the data.

To get an idea of the relative importance of the variables in the model and their influence on the misclassification rate, I ran a forward selection. In each step classification tables were set up based on leave-one-out crossvalidation. The cutpoint was again chosen such that sensitivity equals specificity, and both misclassification rate and area under the ROC curve were computed. A plot of misclassification rate versus the number of variables in the model is given in Figure 5.1. It shows that considerably smaller models with about 20 variables have very similiar predictive performance as the original model with 28 variables.

Hence I tried to find a way of reducing the number of variables in the model while assuring that discriminatory power is maintained. A sound strategy seemed to be to investigate the effect of removing variables that were entered last in the forward selection procedure. The six effects entered last are the two three-way interactions

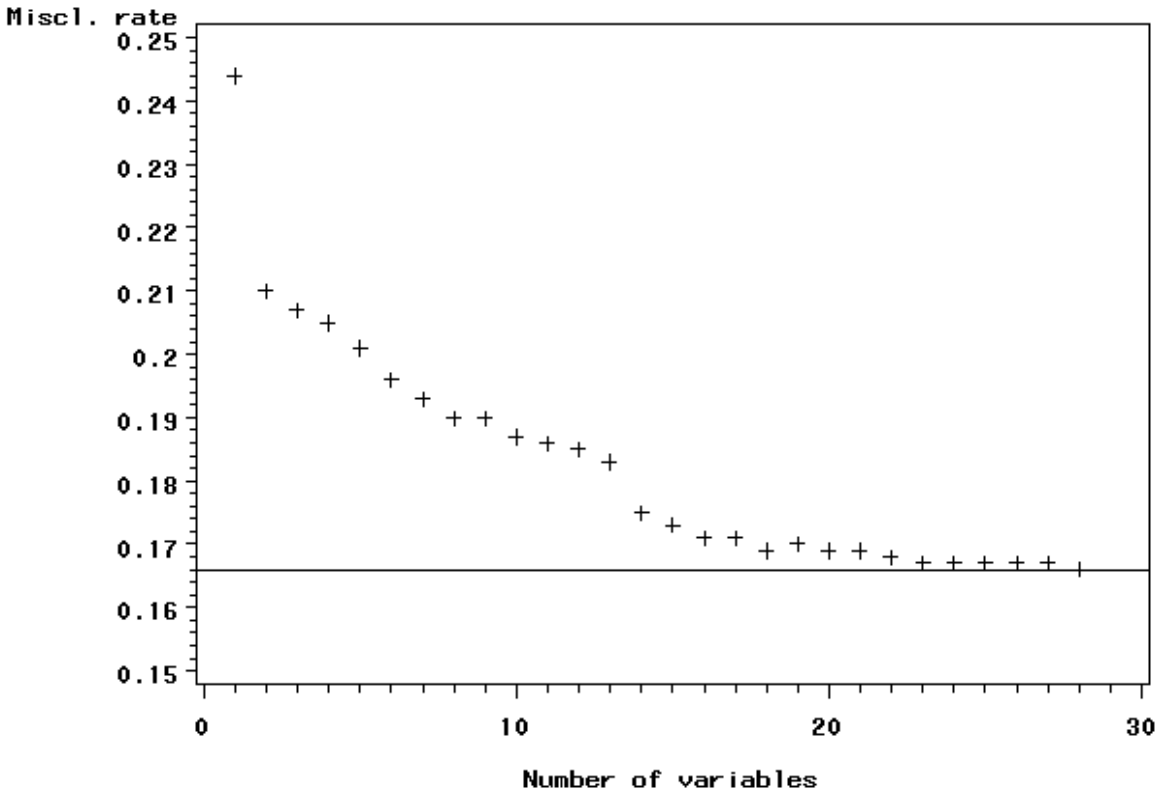

Figure 5.1: Misclassification rate vs. number of variables (model with 28 variables)

(STAERKE*TYPKLH*JTBH and STAERKE*JTBH*KZKHT2), one of the two-way interactions (STAERKE*JTBH) and the main effects ZART, KUNALTER2 and PLZX.

Removing the interaction terms considerably reduces the complexity of the model. ZART (payment option) does not seem to be an important risk factor in the context of the study. The removal of the variable KUNALTER2 (age of policy holder), for which 10% of the values are missing, allows us to use almost all observations (19,981 out of 20,000) for the fitting of the model. Finally, the zip code (categorical variable PLZX with 23 levels) also does not seem to be necessary for achieving the minimal misclassification rate; its removal substantially reduces the number of parameters in the model.

If these 6 variables are removed, then the resulting model has 22 variables and 40 parameters. External validation was carried out in exactly the same way as for the larger model. Table 5.1 shows that both the area under the ROC curve (median of 0.921 for the 10 random partitions) and the misclassification rates (median 16.48%) remain almost unchanged. Hence the smaller model basically has the same predictive performance as the considerably more complex original model.

Table 5.1: Results for the model with 22 variables

Partition	Area under the ROC curve	Misclassification rate (in %)
1	0.923	16.47
2	0.924	16.43
3	0.920	16.49
4	0.920	16.48
5	0.918	16.33
6	0.916	17.05
7	0.923	16.14
8	0.924	15.69
9	0.921	16.57
10	0.917	17.03
median	0.921	16.48
minimum	0.916	15.69
maximum	0.924	17.05

The p-values for the normal approximation to the Pearson chi-square statistic are well above 0.05 for all partitions except one, where it is equal to 0.0356. For this partition, one observation in the validation sample can be identified to have an undue influence on the p-value. It contributes 20% to the value of the Pearson chi-square statistic X^2, and if it is removed for the computation of the test statistic, then the p-value changes to 0.3903 and hence indicates that the model fits. Further investigation has shown that this contract has an unusual combination of characteristics and was not cancelled. The

specific partitioning of the dataset leads to an estimated probability of cancellation of 0.99942 and a large residual.

Moreover, the p-values for Stukel's test also indicate that the model fits. The results for the Hosmer-Lemeshow test are similar to those for the larger model. The p-values depend on the choice of the cutpoints, and at several steps in the modeling process this test has proven not to be a reliable indicator of goodness of fit. Overall, a detailed inspection of overall measures of goodness of fit suggests that the model fits the data despite some problems with the prediction for outliers in the dataset.

Finally, all variables in the model and the way they are entered seem reasonable in the context of the study. Further investigation has shown that removal of additional variables either leads to a significant increase in the misclassification rates and/or results in consistently small p-values for the goodness of fit statistics; especially the interaction terms seem to be important for both goodness of fit and discriminatory power. The best model found that did not include these interactions had a significantly higher misclassification rate and lower values for the area under the ROC curve. On the other hand, it seems that the misclassification rate cannot be reduced by inclusion of more variables or interactions. Altogether this model with 22 variables is a reasonable final model, taking into account prediction error, complexity, goodness of fit and context of information. It will be discussed in more detail in the following section.

Before we come to this discussion, one more comment needs to be made on the influence of outliers on overall measures of goodness of fit like the Pearson chi-square statistic. In the model selection process, a total of 5 observations in the dataset of 20,000

observations were identified to have an excessive influence on the p-values for the normal approximation to the Pearson chi-square statistic.

For example, one of them corresponds to a contract for a 63-year-old vehicle with a vehicle power of 17 kW, which is definitely unusual. The fact that age of the vehicle is treated as linearly increasing on the logit scale resulted in estimated cancellation probabilities above 0.99990. Since the contract was not cancelled, the corresponding Pearson residual was huge and contributed more than 90% to the value of X^2; its contribution to the variance term σ_v^2 was relatively small. The result was a large value for the standardized statistic Z and hence an extremely small p-value (< 0.0001).

Another four influential observations can also be classified as outliers. They are the only contracts with a special type of coverage for motor TPL (DECKAR = 88), and all have an extremely low premium for motor TPL (JTBH = 25.0). Unfortunately, detailed information on the nature of these contracts is not available. All these contracts were cancelled and also had predicted cancellation probabilities greater than 0.9999. Hence the residuals were small, but the contribution to the variance term was large, which lead to an increase in the p-values. If they are ignored, the p-values decrease.

Since the main focus of this study should be all reasonable and common types of contracts in the portfolio, it seems acceptable to ignore these 5 contracts for the logistic regression models.

5.1.3 Discussion of the Final Model

All observations without missing values for any of the variables in the model (19,976 records) were used to fit the model selected in section 5.1.2. It has 40 parameters

and all 22 effects (17 main effects and 5 interactions) are significant, even at a 0.01 level. Based on leave-one-out crossvalidation, the misclassification rate is estimated to be 16.5% (cutpoint 0.55) and the area under the ROC curve is computed as 0.921. These results agree with those obtained using external validation in the previous section.

Since the cutpoint is chosen such that sensitivity equals specificity, we can note that the resulting classification rule classifies 83.5% of the cancelled contracts as cancelled and 83.5% of the not cancelled contracts as not cancelled.

The value of 0.921 for the area under ROC curve has the following interpretation. Among all $9,980 \times 9,996 = 99,760,080$ pairs of cancelled and not cancelled contracts, the cancelled contract has a higher predicted cancellation probability in 92.1% of the cases. A plot of the ROC curve is given in Figure 5.2. The conclusion is that the model provides excellent discrimination between the two outcomes and is capable of predicting the outcome correctly in approximately 83.5% of the cases.

The maximum likelihood estimates of the parameters can then be used to estimate odds ratios as discussed in section 3.6. Some selected estimated odds ratios along with the corresponding 95% confidence intervals based on the Wald statistic are given in Table 5.2; LCL and UCL denote the lower and upper confidence limit.

Table 5.2: Point and interval estimates for selected odds ratios

Variable	Unit	Estimate	LCL	UCL
Years in the portfolio	increase of c=5	0.778	0.755	0.801
Tariff group	Farmers vs. Normal	1.848	1.234	2.769
Tariff group	Officials vs. Normal	0.333	0.297	0.372
Restricted number of drivers	Yes vs. No	0.707	0.619	0.807
Restricted mileage	Yes vs. No	0.503	0.430	0.589
Home owner discount	Yes vs. No	0.289	0.248	0.337
Employee tariff	Yes vs. No	0.309	0.232	0.412

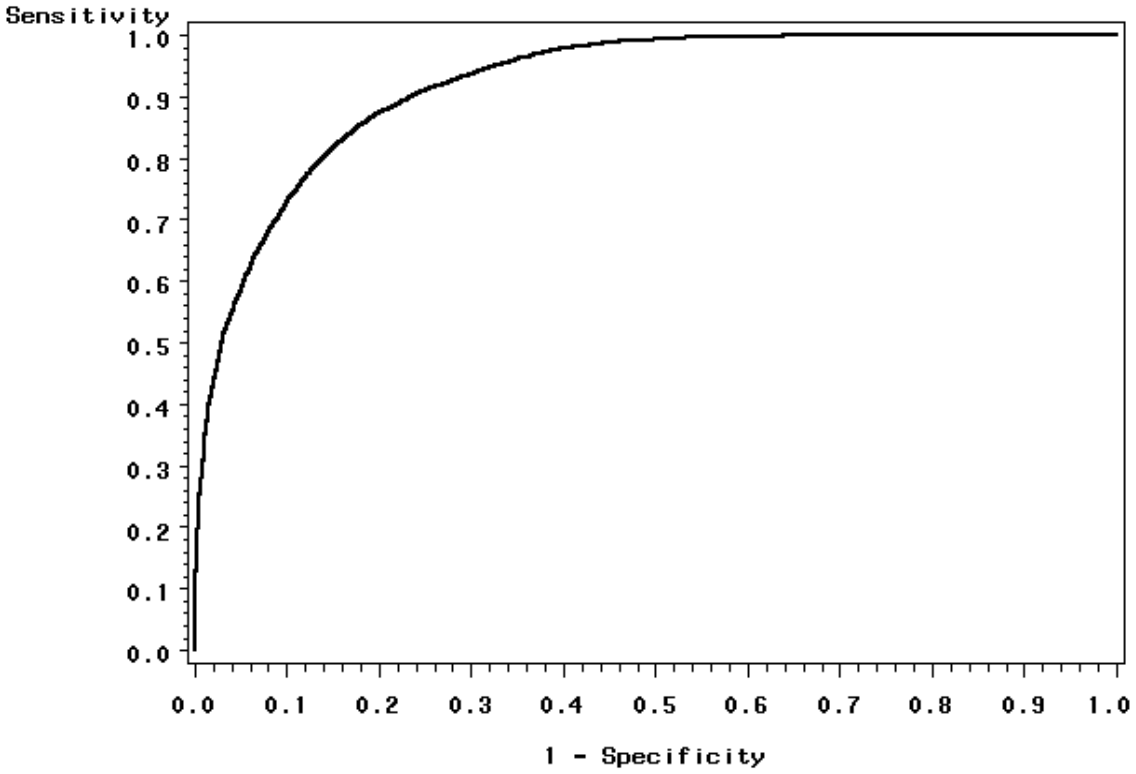

Figure 5.2: ROC curve for the model with 22 variables

Further considerations are necessary if both a linear term and an indicator variable for an effect are used as predictors, as it is the case for the age of the vehicle (FALTER with indicator FALTER0), the percentage rate for motor TPL (BEITSAT3 with indicator BEITSAT2) and the premium for vehicle insurance (JTBF with indicator KZFV). For the definition of these variables see section 5.1.1. As Hosmer and Lemeshow (2000, p. 104) note, this setup allows to estimate two types of odds ratios. For FALTER, for example, both the odds ratio corresponding to a change of c units for vehicles one year or older and the odds ratio comparing contracts for new vehicles to contracts for vehicles of a fixed age greater than zero can be estimated. Point estimates and confidence intervals for these odds ratios are given in Table 5.3. For details on the estimation see section 3.6.

Table 5.3: Odds ratios for FALTER, BEITSAT and JTBF

Variable	Unit	Estimate	LCL	UCL
Age of the vehicle	increase of c=1	1.225	1.209	1.242
Age of the vehicle	new vs. 1 year	5.719	4.405	7.426
Age of the vehicle	new vs. 5 years	2.539	1.994	3.233
Age of the vehicle	new vs. 10 years	0.920	0.729	1.161
Age of the vehicle	new vs. 15 years	0.333	0.261	0.425
Percentage rate TPL	increase of c=10	1.306	1.276	1.335
Percentage rate TPL	30 vs. 100	0.155	0.132	0.181
Percentage rate TPL	50 vs. 100	0.264	0.235	0.295
Percentage rate TPL	75 vs. 100	0.514	0.485	0.543
Percentage rate TPL	125 vs. 100	1.948	1.840	2.061
Premium vehicle insurance	increase of c=250	1.101	1.080	1.123
Premium vehicle insurance	250 vs. 0	0.752	0.670	0.845
Premium vehicle insurance	500 vs. 0	0.829	0.738	0.930
Premium vehicle insurance	1000 vs. 0	1.004	0.886	1.139
Premium vehicle insurance	1500 vs. 0	1.218	1.052	1.410
Premium vehicle insurance	2000 vs. 0	1.477	1.242	1.755
Premium vehicle insurance	3000 vs. 0	2.170	1.712	2.752

All these odds ratios should be interpreted with some care. They are good indicators for the direction of the association, but their actual numerical value may be misleading if there are strong associations among the predictor variables in the model. For this model for example, the odds ratios for FANZ3 (restricted number of drivers), KZWF3 (restricted mileage) and KZEHB1 (home owner discount) may be misleading since possible values for these variables depend on the tariff generation (variable KZKHT2). Hence they estimate the odds of cancellation for customers who get the corresponding discount compared to those who either are not offered the discount or do not qualify.

Associations between the variables STAERKE (vehicle power), TYPKLH (type class motor TPL), JTBH (annual premium motor TPL) and KZKHT2 (tariff generation motor TPL) resulted in the inclusion of five two-way interactions between these variables

in the selected model. First of all, a fixed DM premium increase seems to have different impacts on the cancellation behavior depending on the type of the vehicle as described by vehicle power and type class (interactions JTBH*STAERKE and JTBH*TYPKLH). Furthermore the effects of premium, vehicle power and type class on the cancellation behavior seem to depend on the tariff generation (interactions JTBH*KZKHT2, STAERKE*KZKHT2 and TYPKLH*KZKHT2).

As discussed in section 1.2, the vehicle power is one of the three tariff factors for the oldest tariff generation; for the three newer generations the type class is used instead. Thus, the level of association between the premium and both vehicle power and type class depends on the tariff generation. Overall, the effect of the premium on the cancellation behavior does not seem to be as simple as described by a linear or quadratic model. Because of this complex structure, estimation of odds ratios for any of these variables is complicated and does not add to the understanding.

Finally, the order in which variables are entered in the forward selection procedure can be used as an indicator for the relative importance of the variables. Overall, results can be summarized as follows.

The most important factor seems to be the tariff generation (KZKHT2); using it alone allows us to classify 75% of the contracts correctly. This is partly due to a flaw of the provided dataset. Tariff generation C was introduced in July 1998 and the sample of cancelled contracts was taken from the years 1996-1998. Since the not cancelled contracts were sampled from the total portfolio in 1999, almost all contracts in tariff generation C in the dataset are not cancelled. This is another reason why results have to be interpreted with care.

Other important covariates are the number of years in the portfolio (VWDAU) and the age of the insured vehicle (FALTER). The longer contracts are in the portfolio, the smaller are their odds of cancellation. In general, the odds of cancellation increase with the age of the vehicle. New vehicles are an exception: for them the odds of cancellation are higher compared to the odds for 1 to 5 year old vehicles.

Also the main tariff factors percentage rate (BEITSAT3, BEITSAT2), tariff group (TGR2), regional classification (REGIO), vehicle power (STAERKE), type class (TYPKLH) and the final premium for motor TPL (JTBH) have considerable influence on the cancellation behavior. For example, the odds increase as the percentage rate increases; especially contracts with percentage rate above 125 are very likely to be cancelled. Moreover, there are significant differences between the tariff groups. The odds of cancellation are higher for farmers and lower for officials, compared to customers in the tariff group "normal".

Variables representing various discounts (KZWF3, FANZ3, KZEHB1 and VERBU1) that may apply to a contract are also used as predictors in the model, but they are of minor importance. For all these types of discounts, it can be seen that the odds of cancellation are smaller for customers who get the discount. Part of this effect is due to the fact that some of the discounts were not offered in older tariff generations.

The only variable in the model related to vehicle insurance is the premium for vehicle insurance (JTBF with indicator KZFV). The odds of cancellation are higher for customers who have not included vehicle insurance compared to those who pay relatively small premiums for vehicle insurance. On the other hand, the odds of cancellation are higher for contracts with a large premium for vehicle insurance, compared to those

without vehicle insurance. Neither type class nor percentage rate needs to be used as a predictor in order to get an accurate classification rule. It seems that motor TPL is the component of the contract that has the largest influence on the cancellation behavior. More detailed information on the type of coverage for vehicle insurance (partial or full) may be helpful for further improvements of the classification rule.

Also the age of the policy holder does not seem important and is not used as a predictor. However, factors likely to be related to the age, for example, the number of years in the portfolio and the percentage rate, are used as predictors and may account for differences in the cancellation behavior between different age groups.

5.2 Classification Trees

5.2.1 Variable Screening

Contrary to logistic regression models, the complete dataset consisting of 20,000 records can be used for building classification trees. Missing values are handled as discussed in chapter 4, and classification trees are robust with respect to outliers. For the definition of the set of possible splits we have to distinguish between continuous and categorical predictors.

The results of the variable screening for logistic regression models discussed in section 5.1.1 are useful at this point. Variables that were entered with linear or quadratic terms in the logistic regression models are now treated as continuous. Also categorical variables are used in the same way as for logistic regression models. It is not necessary to

combine categories with small counts, since there are no numerical problems associated with that.

Hence the following 14 variables are treated as continuous for the tree models: VWDAU, KUNALTER2, ZART, FALTER, STAERKE, KZWF2, TYPKLH, TARKLA2, BEITSAT, JTBH, TYPKL, TARKLAFV, FELD68 and JTBF. It should be noted that the value zero for TYPKLH means that the type class for motor TPL is not available because the contract is based on an old tariff generation (KZKHT = O, 5 or 6). For all variables related to vehicle insurance (TYPKL, TARKLAFV, FELD68 and JTBF), a value of zero means that vehicle insurance is not included in the contract.

Additionally, I used 17 categorical variables (number of levels in parentheses): AUSLK1 (2), PLZX (23), REGBE (9), KZKHT (6), DECKAR2 (3), REGIO (10), TGR2 (4), FANZ1 (7), KZEHB1 (2), KZFV (2), SORAB1 (2), VERBU1 (2), TARKLA_0 (2), TARKLA_12 (2), TARKLA_S (2), TARKLA_M (2), TARKLAFV_0 (2) and TARKLAFV_12 (2). For details on these variables see section 5.1.1.

Quadratic terms, interactions and indicator variables like FALTER0, BEITSAT2 and KZFV are not included, since the tree procedure automatically detects nonlinear structures and interactions. Altogether 31 variables are used for growing classification trees.

5.2.2 Model Selection and Model Assessment

The Enterprise Miner package for SAS was used to fit tree models. Overall, the SAS Enterprise Miner can be used to build classification trees according to the established methodology discussed in chapter 4. Only cost-complexity pruning is not available;

instead the subtree with the smallest misclassification rate on the validation sample is found for every possible number of leaves. Details of this implementation are discussed in appendix D.

The dataset was divided into a training sample (2/3) used to grow and prune the tree and a validation sample (1/3) used to select the subtree with smallest misclassification rate. As for logistic regression models I used 10 random partitions to confirm the results.

Preliminary investigation showed that the best strategy is to grow the largest possible tree first and then prune it as described in chapter 4. That is, nodes are split until they are either pure or contain only one observation. Alternatively, if splitting is stopped as soon as the number of observations in a node is below some threshold N_{min}, then the computational effort is reduced since the resulting tree is smaller. The default in the Enterprise Miner package, for example, is the number of observations in the training sample divided by 100. But after the pruning process, the resulting trees have a higher misclassification rate for a given number of leaves than trees that are grown to maximal size first. Hence this is not an advisable strategy, at least if there are no computational problems associated with the growth of the largest possible tree.

Another option for the tree growing process is the choice of the impurity measure (Gini index or entropy). Both impurity measures result in structurally similar trees for a given partition. Table 5.4 contains minimum misclassification rates on the validation sample and corresponding number of leaves for both Gini index and entropy measure, for the 10 random partitions of the dataset.

Table 5.4: Comparison of Gini index and entropy measure

Partition	Gini index		Entropy	
	Number of leaves	Miscl. rate	Number of leaves	Miscl. rate
1	345	15.04%	300	15.70%
2	326	14.92%	328	14.92%
3	289	15.33%	326	15.42%
4	257	15.39%	342	15.27%
5	253	15.22%	310	15.34%
6	261	15.84%	316	16.53%
7	300	14.43%	325	15.07%
8	328	14.08%	347	15.46%
9	364	14.79%	315	15.46%
10	310	15.09%	336	15.48%
median	305	15.07%	325	15.44%
minimum	253	14.08%	300	14.92%
maximum	364	15.84%	347	16.53%

For all partitions except one, the minimum misclassification rate for trees grown using the Gini index is less than or equal to the misclassification rate for the corresponding tree based on the entropy measure. The median of the differences is 0.51%. Moreover, the trees based on the Gini index are smaller on average than those based on the entropy measure. Further calculations (not shown here) indicated that the trees based on the Gini index for which the misclassification rate is equal to the minimum misclassification rate for the corresponding tree based on the entropy measure are consistently smaller (median: 147 leaves). Hence the Gini index seems preferable for this particular dataset and will be used to grow classification trees.

To illustrate the pruning process and motivate the choice of the appropriate subtree, it is useful to consider a plot of the misclassification rate on both training and

validation sample versus the number of leaves of the subtree. The corresponding plot for random partition #10 from Table 5.4 for the Gini index is given in Figure 5.3.

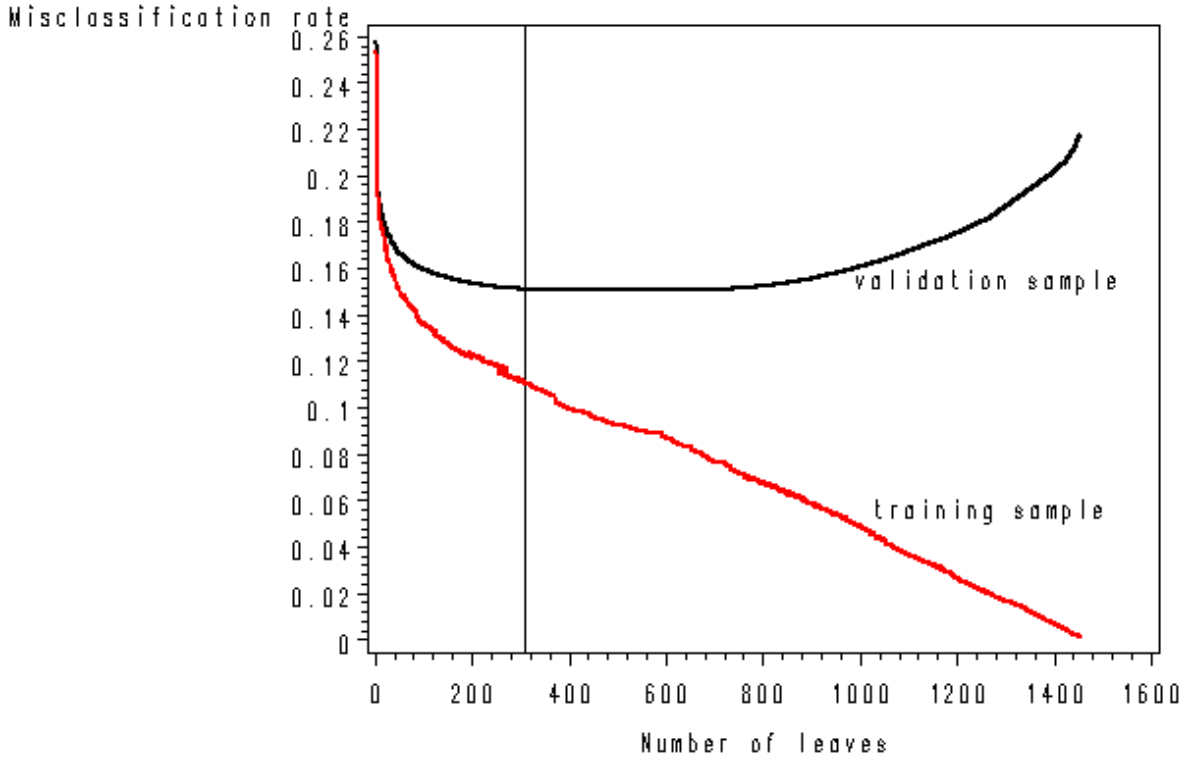

Figure 5.3: Misclassification rates vs. number of leaves (partition #10, Gini index)

For the training sample, we see a rapid decrease for the first about 100 leaves, followed by a linear decrease for larger trees, with some minor fluctuations. For the largest tree with 1,453 leaves, the misclassification rate is 0.14%.

The curve for the validation sample looks entirely different. For the first 40 leaves, the misclassification rates are very similar. From then on, the mislassification rate on the validation sample is greater than on the training sample, and the difference is increasing. This illustrates the fact that estimation of model performance based on the training sample can give overly optimistic results. However, the misclassification rate on the validation sample decreases until it reaches its minimum of 15.09% for 310 leaves (vertical

reference line in Figure 5.3). Trees between 300 and 800 leaves seem to have very similar misclassification rates on the validation sample. But there is a significant increase in the misclassification rate for subtrees with more than 800 leaves. This is a consequence of the fact that these large trees overfit the data. The largest possible tree with 1,453 leaves has a misclassification rate of 21.72% on the validation sample.

Since the pruning is based on the misclassification rate on the validation sample as discussed in chapter 4, it is reasonable to select the subtree with the smallest number of leaves such that its misclassification rate is minimal. However, the plot suggests that a misclassification rate on the validation sample very "close" to the minimum can be achieved using a substantially smaller tree. As mentioned in chapter 4, one suggestion is to choose the smallest subtree such that the misclassification rate on the validation sample is within one estimated standard error of the minimum ("one standard error rule"). Application of this rule leads to the results given in Table 5.5 (for the Gini index). The median of the misclassification rates has increased by 0.43%, but also the number of leaves has decreased significantly; on average, it was reduced by 45%.

However, the resulting trees still have between 131 and 240 leaves, which limits the interpretability. If such a tree is actually used for predicting the cancellation behavior for the complete portfolio of the insurance company, application of the "one standard error rule" may substantially reduce computing time, where the estimated predictive performance is only slightly below the optimum. This illustrates that there needs to be a trade-off between predictive performance and complexity which depends on the context of the problem and available computer resources. For a detailed discussion in the following section I will consider the trees selected according to the "one standard error rule".

Table 5.5: Minimum misclassification rate and "one standard error rule"

Partition	Minimum		One standard error rule	
	Number of leaves	Miscl. rate	Number of leaves	Miscl. rate
1	345	15.04%	192	15.48%
2	326	14.92%	167	15.36%
3	289	15.33%	138	15.77%
4	257	15.39%	133	15.83%
5	253	15.22%	148	15.66%
6	261	15.84%	131	16.29%
7	300	14.43%	175	14.86%
8	328	14.08%	196	14.51%
9	364	14.79%	240	15.22%
10	310	15.09%	163	15.53%
median	305	15.07%	165	15.50%
minimum	253	14.08%	131	14.51%
maximum	364	15.84%	240	16.29%

5.2.3 Discussion of the Selected Classification Trees

For the tree models selected in the previous section, we can crossclassify observed and predicted outcomes and analyze the resulting 2×2 tables. In addition to the misclassification rates given in Table 5.5, sensitivity and specificity are calculated. For all partitions, the sensitivity (approximately 86%) is slightly higher than the specificity (approximately 84%). This means that cancelled contracts are predicted slightly more accurately than not cancelled contracts. Overall, approximately 85% of the contracts are classified correctly.

The large size of the trees limits the interpretability of the tree structure. Some insight into the relative importance of the 31 predictor variables can be gained by examination of the importance rankings. These rankings are quite similar for the 10 partitions; the ranking for partition #10 along with the number of times a variable is used for splitting is given in Table 5.6.

Table 5.6: Importance ranking (partition #10, 163 leaves)

Variable	Label	Importance	Rules
KZKHT	tariff generation TPL	1.0000	7
JTBH	premium TPL	0.8745	22
FANZ1	number of drivers	0.8544	1
KZWF2	mileage class	0.8454	1
KZEHB1	home owner discount?	0.8427	3
TARKLA2	no-claims class TPL	0.7256	3
VWDAU	years in portfolio	0.6983	15
DECKAR2	type of coverage TPL	0.6957	0
BEITSAT	percentage rate TPL	0.6669	14
TYPKLH	type class TPL	0.6661	2
REGIO	regional classification	0.6514	18
FALTER	age of vehicle	0.6168	12
PLZX	zip code area	0.5334	26
JTBF	premium vehicle insurance	0.5212	8
TARKLA_12	SF 1/2 TPL?	0.5118	0
STAERKE	vehicle power (kW)	0.4969	9
FELD68	percentage rate vehicle insurance	0.4635	1
TARKLAFV	no-claims class vehicle insurance	0.4506	2
TYPKL	type class vehicle insurance	0.4442	2
REGBE	county	0.4380	1
TARKLA_0	class 0 TPL?	0.4287	0
KUNALTER2	age of policy holder	0.3000	12
TARKLAFV_12	SF 1/2 vehicle insurance?	0.2156	0
AUSLK1	foreigner?	0.1883	0
TARKLA_S	class S TPL?	0.1833	0
TARKLAFV_0	class 0 vehicle insurance	0.1817	0
VERBU1	employee tariff?	0.1551	0
TGR2	tariff group TPL	0.1297	2
ZART	payment option	0.1213	0
SORAB1	special discount?	0.1180	1
TARKLA_M	class M TPL?	0.0000	0

This ranking indicates that tariff generation and premium for motor TPL are the most useful variables for distinguishing between cancelled and not cancelled contracts. The high ranking for the tariff generation is partly due to the way the sample was taken (see discussion later in this section).

Furthermore, some of the tariff factors for TPL insurance play an important role for the recursive partitioning of the covariate space. And, of course, the number of years in the portfolio seems to be one of the main factors influencing the cancellation behavior. All variables related to vehicle insurance play a minor role in the tree construction process.

Among the least important variables are the indicator variables for claims classes in both motor TPL and vehicle insurance, and variables representing special discount, payment option, tariff group, employee tariff and nationality. Also the age of the policy holder is ranked low, even though it is used for 12 splits. This may partly be due to the fact that it has a lot of missing values, and they count against its importance. On the other hand, other variables that are likely to be related to the age (for example, the number of years in the portfolio and the number of years without a claim) may account for differences in cancellation behavior between age groups.

Such a ranking only gives an idea about which variables are useful for the successive partitioning of the covariate space; it does not give any information about how the actual partition is achieved. A tree ring display like the one in Figure 5.4 can be used to visualize the partition defined by a classification tree. For concept and definition of a tree ring see appendix D. For this example, the coloring of the tree ring indicates the percentage of observations from the training sample that are classified correctly (see legend). The overall misclassification rate on the training sample is 12.55%.

Figure 5.4: Tree ring (partition #10, 163 leaves)

First of all, the tree ring shows that the partition is highly complex, as expected for a tree with 163 leaves. The covariate space is partitioned into many small subsets, but there are also several larger subsets. Moreover, the tree ring is unbalanced. In some areas, the final partition is defined by only a few splits, whereas in others up to 20 splits are necessary to define a leaf.

Overall, the coloring reflects the good discriminatory power of the corresponding classification rule. Most leaves have a misclassification rate below 30%; especially some large leaves classify observations with more than 90% accuracy. Only for a few leaves the black color indicates a misclassification rate above 30% for the training sample.

Further insight into the tree structure and the definition of the largest leaves can be gained by examination of the first few levels of the classification tree. The first three levels are given in Figure 5.5.

First, the training set is split into contracts based on the newest tariff generation C and contracts in all other tariff generations (variable KZKHT). Among the contracts in tariff generation C only 2.4% are cancelled, a fact that is mostly due to the way the sample was taken. Tariff generation C was introduced in July 1998, and the sample of cancelled contracts is taken from all cancelled contracts in the years 1996 - 1998. The not cancelled contracts, on the other hand, were sampled from the total portfolio in 1999.

Competing splits on the root node try to mimic the chosen split. For example, a split on the variable KZWF2 (mileage classes 1 through 4 vs. mileage class 5) results in a similar decrease in impurity and assigns almost 90% of the observations to the same branch as the primary split does. This however is due to the fact that classes 2 through 4 are only offered in tariff generation C, and class 1 only in tariff generations A, B and C.

All contracts based on older tariffs (45%) have a value of 5 for KZWF2. This discussion illustrates that a more balanced sampling method would result in structurally different trees. Hence all interpretations must be handled with care, since they may be biased because of the way the sample was taken.

Among the contracts based on tariff generation C, those with a percentage rate less than 110 (variable BEITSAT) are very unlikely to be cancelled. Leaf #4 with 3,221 observations corresponds to the largest white area of the tree ring in Figure 5.4. If the percentage rate is above 110, but the contract has been in the portfolio for at least one year, then it is unlikely to be cancelled (leaf #11 with 70 observations). On the other hand, one more split based on the premium for vehicle insurance is used to classify the remaining 41 contracts that are new in the portfolio and have a percentage rate above 110 (node #10). Overall, even for the relatively homogeneous subset of tariff generation C contracts, information on the percentage rate for TPL insurance, the number of years in the portfolio, and the premium for vehicle insurance can be used to improve the accuracy of the prediction rule.

Node #3, which contains all contracts not in the newest tariff generation, is split using the variable TYPKLH (type class motor TPL). The left branch corresponds to all contracts with a value of 0, 10 or 11 for TYPKLH. That is, since the number of contracts with type class 10 or 11 is small and since a value of 0 means that the contract is in one of the three oldest tariff generations (5, 6 or O), we basically have a second split based on the tariff generation.

For both branches, splits depending on the number of years in the portfolio result in further reduction of the overall tree impurity (see nodes #12/13 and #14/15). In the

left branch (node #6), for example, contracts that have been in the portfolio of the VKB for at most 3 years form a relatively pure node (#12). Two more splits, again based on the number of years in the portfolio (VWDAU) and then on the tariff generation (KZKHT), result in a leaf with 1,146 observations and a misclassification rate of 0.2%. This leaf corresponds to the area labeled #47 in Figure 5.4. Also the other white area of the tree ring (labeled #49) is defined by two more splits of node #12 based on tariff generation and number of years in the portfolio.

The set of contracts with a value of at least 4 for the number of years in the portfolio (node #13), on the other hand, is divided in the next step in longtime contracts (25 years or more) and contracts between 4 and 24 years in the portfolio (nodes #26 and #27). The latter subset is then split according to the age of the vehicle and the premium for motor TPL. For 624 contracts, the vehicle is older than 13 years and the premium is below DM 1370.35; they form leaf #86 in Figure 5.4 (89.4% cancelled). If the vehicle is at most 13 years old, then 6 more splits are necessary to define leaf #566 with 718 observations, among which 79.8% are cancelled.

Now let us consider node #7 which contains all contracts in tariff generation A and B with a type class of 12 or higher and is balanced between cancelled and not cancelled contracts. Two splits based on VWDAU define leaf #28 that contains all 481 new contracts (VWDAU = 0) in this subset; they are all cancelled. Also the customers with exactly one year in the portfolio are likely to cancel if their percentage rate is above 110 (leaf #55: 147 contracts, 98.6% cancelled).

On the other hand, among the contracts that have been in the portfolio for at least two years (node #15), the majority is not cancelled. Especially those for which

the percentage rate is small and the vehicle is relatively new are unlikely to be cancelled (leaf #96: 246 contracts, 91.9% not cancelled). For all others, the prediction is not very accurate, as the predominantly grey and black color of the tree ring in Figure 5.4 shows. A larger tree, for example the one with 310 leaves and minimum misclassification rate, is capable of predicting the cancellation behavior of this subset of contracts more accurately.

Overall, in the first 6 levels of this classification tree, almost all primary splits are based on the following 4 variables: KZKHT, VWDAU, BEITSAT and FALTER. Two splits are based on PLZX (zip code area) and one on the premium for vehicle insurance; the split based on TYPKLH in node #3 is basically also a split based on the tariff generation (see discussion above).

Although splits in lower levels of the tree operate on subsets of the complete training sample, some common features of the splits on these four variables can be noticed. New customers seem to be very likely to cancel (VWDAU $= 0$); also the first few years in the portfolio are critical. Longtime customers (VWDAU ≥ 25) are unlikely to cancel. Moreover, a high percentage rate for motor TPL insurance (BEITSAT ≥ 110) is a good indicator for cancellation. On the other hand, a very small percentage rate indicates that the customer is unlikely to cancel. In addition, contracts for old vehicles are likely to be cancelled.

The overall conclusion is that the complexity of the tree structure limits the interpretability. However it is possible to identify some key variables used to build the upper levels on the tree. Classification trees definitely provide accurate classification rules for this problem.

CHAPTER 6

COMPARISONS AND CONCLUSION

6.1 Comparison of Logistic Regression Models and Classification Trees

In the first section of this final chapter I would like to compare the two modeling approaches used in this thesis for the prediction of cancellation behavior. First, assumptions, strengths and weaknesses of both logistic regression models and classification trees are discussed in general. Breiman et al. (1984, chap. 2) mention some of the advantages of the tree structured approach; Steinberg and Cardell 1998) compare the two approaches in detail. Additionally, the results for the provided dataset are compared.

6.1.1 General Comparison

First of all, we should recall the two basic assumptions for the logistic regression model: the binary outcome is modeled as Bernoulli distributed with a probability of success that depends on the covariate vector, and the logit of the probability of success is modeled via a linear predictor (see section 3.1). That is, a functional form is required, and we assume a linear or curvilinear data structure. This results in smooth, continuous predicted probabilities. A small change in a continuous variable x results in a small change in the predicted probability y.

Classification trees, on the other hand, do not require specification of a functional form and are nonparametric procedures. Only the set of possible splits, the rule for selecting the best split, and the criterion for selection of the best tree need to be specified before the classification tree can be grown. One consequence is that predicted probabilities are discontinuous: a small change in a continuous predictor x may result in a large change in the predicted probability y. Classification trees can be applied in a natural way to problems with polychotomous outcomes (more than $K = 2$ classes). Generalizations also exist for logistic regression models (see Hosmer and Lemeshow (2000), chap. 8).

As Steinberg and Cardell (1998) note, logistic regression requires "hand-built" models. The variable screening is an essential part of the modeling process, since we need to assure that the assumption of linearity on the logit scale is satisfied. If it is not, transformations of continuous predictors (for example, polynomial terms, log, square-root) should be considered. Interactions are hard to detect, if they are not obvious from the context of the study. This presents a problem if the type of study is new and there are no results from previous, similar studies that could be used as a guideline. Overall, the model-building process is time-consuming and requires expert knowledge in order to give satisfactory and reliable results.

Classification trees automatically seperate irrelevant from relevant predictors, and transformations do not need to be considered. But of course, performance can be enhanced by a careful selection of predictor variables. Furthermore, the tree procedure automatically detects interaction structures and nonlinear relationships and uses them in the growing process. Overall, tree-based methods are faster to apply since they are highly automated, and they are easier to use, especially for non-statisticians.

Additionally, the methodology for classification trees as discussed in chapter 4 provides a natural way of handling missing values via surrogate rules (see section 4.5). For logistic regression, two options for the handling of missing values are imputation and deletion. Imputation is complicated and time-consuming, especially for large databases. If observations with missing values are deleted, then the sample size is reduced, which may also present a problem.

Another advantage of tree models is that they are not affected by outliers or errors in the dataset and hence can handle "dirty" data. Every data point has weight 1 in the tree growing process and the resulting classification rules have a robustness property similar to the median. Contrary to that, logistic regression models are sensitive to outliers because of the specified functional form.

Altogether, the main advantage of classification trees is that they can handle complex data structures (many predictors, large number of records) in a variety of contexts. They can also easily detect and use local structures in large databases. The structure is graphically displayed by a tree diagram. This makes them a very flexible data analysis tool.

The possibility of uncovering structure of the data by a classification tree is limited by the instability of the tree structure. There may be nodes where splits on several variables give almost the same decrease in impurity. Since the data are "noisy" and the partition in training and validation sample is random, the choice of the split at such a node may depend on the partition. If different splits are chosen for different partitions, then the evolution of the tree from that node downward will differ (see Breiman et al. (1984, p. 156)). Sometimes inspection of competing splits at a node may help in the

interpretation of the tree structure. The main reason for this instability is the hierarchical structure of a tree. The variation in the prediction may be reduced by the use of Bagging, that is, by averaging over several trees based on different bootstrap samples (see Hastie et al. (2001, p. 246)).

Logistic regression models, on the other hand, can effectively capture the global features of the data. The output of logistic regression models describes this structure and gives the possibility to interpret the results in terms of odds ratios. Also many non-linear structures can be reasonably approximated with a linear structure. Hence, even an incorrectly specified logistic regression model can perform fairly well.

Classification trees, however, are weak at capturing linear structure. The structure is recognized, but it cannot be represented efficiently and it is not obvious from the output. Sometimes a large tree is necessary to represent a fairly simple linear structure. This is a common feature of nonparametric procedures: whenever the assumptions of the corresponding parametric methods are satisfied, the parametric model seems to perform better.

Application of logistic regression models becomes complicated and time-consuming as the data structure becomes more complex. Local structure in large datasets is in general hard to detect by a logistic regression model. Overall, both classification trees and logistic regression models are powerful data analysis tools and have demonstrated remarkable accuracy in a variety of problems.

6.1.2 Results for the Cancellation Prophylaxis Study

The first result to be mentioned here is that the model building process was significantly faster and also easier for classification trees. Only a few options needed to be set, and classification trees were built in a rather straightforward way.

Due to the large number of possible covariates, the model selection process for logistic regression models was sophisticated and time-consuming. For every variable, the appropriate functional form had to be found and stepwise procedures were used to seperate important from irrelevant predictors. The assessment of goodness of fit was further complicated by the undue influence of outliers on the test statistics (see discussion in section 5.1.2). Furthermore, the drawbacks of the Hosmer-Lemeshow test became apparent in the analysis of this dataset.

Missing values for the variable representing the age of the policy holder resulted in a decrease of the sample size of about 10%. The final model, however, did not include this variable and was therefore fitted using almost all observations.

As far as the predictive performance is concerned, both modeling approaches yield similar results as the plot of ROC curves in Figure 6.1 shows. Misclassification rates for classification trees were slightly below those for logistic regression models: The difference was 1 to 1.5 percentage points, depending on whether the "one standard error rule" was applied or not. This means that the most accurate logistic regression model had an average misclassification rate about 10% higher than the most accurate classification trees. It seems as if there is a lot of linear structure in these data, and logistic regression models are capable of effectively representing it. The smaller misclassification rate for trees is probably due to the fact that they additionally recognize local structure.

Figure 6.1: Comparison of ROC curves (partition #1)

For both approaches, the resulting models are quite complex: the final logistic regression model contains several interaction terms and a total of 40 parameters. Point and interval estimates for odds ratios give an idea of the actual influence of the variables on the cancellation behavior. This allows identification of risk factors for cancellation and quantification of their influence. However, assocations among predictor variables limit the interpretability.

The large number of leaves for the selected classification trees limits the comprehensibility of the tree structure. Additionally, the structure of the tree depends on the partition; the upper levels are ususally almost identical, whereas there are significant differences in the lower levels of the tree. But also for these models the most important factors for the cancellation behavior have been identified. The complexity of the tree

structure also indicates that the tree tries to represent linear structure in these data, but cannot do this effectively. For every classification tree, we can set up a corresponding logistic regression model by using indicator variables for the leaves as predictors. Since the number of parameters of such a logistic regression model is equal to the number of leaves of the tree, we see that for this study classification trees are more complex than the selected logistic regression model (with 40 parameters). However, in order to achieve a misclassification rate close to the misclassification rate of the final logistic regression model, we only need a classification tree with about 50 leaves.

Overall, both types of statistical models discussed here could be used for the prediction of cancellation behavior. Logistic regression models seem to be easier to interpret, whereas classification trees are slightly more accurate. The choice of the method should also take into account the time factor, especially if the costs of the study are a major concern.

6.2 Suggestions for Further Study

6.2.1 Alternative Modeling Approaches

A variety of methods has been proposed for classification problems like the one discussed in this thesis. Hastie et al. (2001) give an overview over a wide range of modeling approaches, including classical and modern statistical algorithms, tree-based models and neural networks. Lim, Loh, and Shih (2000) have compared 33 classification algorithms with respect to accuracy, complexity and training time, using 32 sample datasets.

Four algorithms show consistently good performance, with respect to both the mean misclassification rate and the rank of the mean misclassification rate. A spline-based statistical algorithm called POLYCLASS is placed at the top; however, it requires relatively long training time. Logistic regression ranks second and, hence, higher than all tree-based methods. The most accurate tree algorithm is QUEST (Quick Unbiased and Efficient Statistical Tree). It is quite similar to the CART algorithm described in Breiman et al. (1984), which was applied in this thesis. The exhaustive split search algorithm used in the CART methodology is biased towards selecting variables that afford more splits. QUEST uses a different split selection criterion, which tries to reduce the bias in variable selection. For details on QUEST, see Loh and Shih (1997).

The tree algorithms CART, QUEST and C4.5 (see Quinlan (1993)) had the best combination of misclassification rate and speed in this study. However, C4.5 tends to produce trees with twice as many leaves as those from CART and QUEST.

Finally, the classical statistical method of linear discriminant analysis is also ranked among the best classification algorithms in this study. This is surprising, since its key assumption (multivariate Gaussian distribution for the class densities) is violated if we have a mixture of categorical and continuous predictors. In addition, it is fast and widely available. Similarly, Hastie et al. (2001, p. 105) point out that logistic regression models and linear or quadratic discriminant analysis models often give very similar results, even if discriminant analysis is used inappropriately with qualitative predictors. However, logistic regression models seem to be safer and more robust in such a situation since they rely on fewer assumptions.

Hence, newer tree-based methods like QUEST (which also incorporates a variant of linear discriminant analysis) may be reasonable alternatives to the selected modeling approaches. Overall, logistic regression models seem to be competitive with tree models if applied correctly. In addition, there have been attempts to combine logistic regression models and classification trees to form a hybrid model (see Steinberg and Cardell (1998)).

6.2.2 Improvement of the Cancellation Prophylaxis Study

Experiences during the modeling process for this cancellation prophylaxis study lead to the following suggestions for possible improvements of the study.

First, we should make sure that all available characteristics of a contract are actually recorded. In the provided dataset, for example, information on the type of coverage in vehicle insurance and on deductibles in both motor TPL and vehicle insurance was missing. Additional information may result in further improvement of the predictive performance of the models. More detailed information on the meaning of the covariates could also assure that unusual contracts like those discussed in section 5.1.2 are excluded from the study in advance.

If a case-control study design is chosen, as for the provided dataset, then the sampling method should assure that cases and controls are sampled from the same time period and not, as was the case here, from separate time periods. This would lead to more unbiased results.

We should note again that a case-control study design cannot provide estimated cancellation probabilities for the contracts in the total portfolio of the insurance company; it is only capable of classifying contracts as cancelled or not cancelled. If

the company is interested in estimation of the actual probability of cancellation, a cohort study would be the appropriate approach (see section 2.3). Both logistic regression models and classification trees can be applied to data from a cohort study in exactly the same way as presented in this thesis for a case-control study.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Asmus, W., & Sonnenburg, V. (1998). *Kraftfahrtversicherung* [Motor Insurance]. Wiesbaden, Germany: Gabler.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Cody, R.P., & Smith, J.K. (1997). *Applied Statistics and the SAS Programming Language* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Cox, D.R., & Snell, E.J. (1989). *Analysis of Binary Data* (2nd ed.). New York: Chapman and Hall.

GDV (2001). *2001 Yearbook: The German Insurance Industry*. Berlin, Germany: Gesamtverband der deutschen Versicherungswirtschaft.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Hosmer, D.W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine*, 16, 965-980.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, 119-127.

Lim, T.S., Loh, W.Y., & Shih, Y.S. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning*, 40, 203-229.

Loh, W.Y., & Shih, Y.S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*, 7, 815-840.

Johansson, L. (1999). *Verfahren der Diskriminanzanalyse angewandt auf Daten der Kfz-Versicherung* [Discriminant Analysis Methods applied to Motor Insurance Data]. Unpublished diploma thesis, University of Ulm, Ulm, Germany.

Osius, G., & Rojek, D. (1992). Normal Goodness-of-Fit Tests for Multinomial Models with Large Degrees of Freedom. *Journal of the American Statistical Association*, 87, 1145-1152.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge, Great Britain: Cambridge University Press.

Steinberg, D., & Cardell, N.S. (1998). *The Hybrid CART-Logit Model in Classification and Data Mining.* Retrieved July 23, 2002, from http://www.salford-systems.com/whitepaper.html#hybrid.

Stokes, M.E., Davis, C.S., & Koch, G.G. (2000). *Categorical Data Analysis Using the SAS System.* Cary, NC: SAS Institute.

Stukel, T.A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83, 426-431.

Wirtz, B.W. (2000). *Electronic Business.* Wiesbaden, Germany: Gabler.

Venables, W.N., & Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS* (2nd ed.). New York: Springer.

Versicherungskammer Bayern (2001). *Kfz-Tarife: Vergleich der VKB Typentarife seit 1.7.1998* [Motor Insurance Tariffs: Comparison of the type class based tariffs since 07/01/1998] [unpublished brochure]. Munich, Germany: Versicherungskammer Bayern.

APPENDIX A

VARIABLES IN THE DATASET

VARIABLES IN THE DATASET

This appendix contains an overview of all 40 variables found in the provided dataset; Table A.1 presents label, scale and range for all variables. It also includes the variable KZFV which was not in the original dataset. An inspection of the dataset has shown that it may be useful to define a variable that indicates whether motor vehicle own damage insurance is included in a specific contract or not. This is done by defining KZFV by

$$\text{KZFV} := \begin{cases} 1 & \text{if JTBF} > 0 \\ 0 & \text{if JTBF} = 0 \end{cases}.$$

Variables that do not provide useful information and are therefore not considered during the analysis are put in *italics*. They are discussed in detail below.

The variable VSNR represents the policy number which is used to uniquely identify a contract in the portfolio of the insurance company. The variable VVNR represents a preliminary policy number which is used for administration purposes. Both variables are not relevant for this study.

The variable VERMNR uniquely identifies the responsible insurance agent for the contract. This characteristic may be relevant to check whether an agent is providing good service to his customers. Since no detailed description is available and the number of contracts per insurance agent is very small (there are 4,866 different agents in the sample), I will also ignore this variable. In order to evaluate the performance of insurance agents a different study design must be used.

The date of cancellation for a cancelled contract is stored in the variable STODAT; STODAT1 contains the year of cancellation. The variable BEGDAT1 corresponds to the year in which the first contract with VKB has been concluded. STODAT, STODAT1 and BEGDAT1 will not be used directly for the analysis. However they were used to define the variable VWDAU, representing the number of years the contract was in the portfolio of VKB, by:

$$\text{VWDAU} := \begin{cases} 99 - \text{BEGDAT1} & \text{if STORNO} = 0 \\ \text{STODAT1} - \text{BEGDAT1} & \text{if STORNO} = 1 \end{cases}.$$

The variable GEBJJ represents the year of birth of the policy holder. It is used to find the relevant age of the policy holder by

$$\text{KUNALTER2} := \begin{cases} 99 - \text{GEBJJ} & \text{if STORNO} = 0 \\ \text{STODAT1} - \text{GEBJJ} & \text{if STORNO} = 1 \end{cases}.$$

The original variable KUNALTER, defined by KUNALTER := 99 - GEBJJ does not take into account the fact that the cancelled contracts are sampled from the years 1996 - 1998 and the not cancelled contracts from 1999. Thus GEBJJ and KUNALTER will be ignored; I will use KUNALTER2 instead.

HERSTELL is the number that uniquely identifies the manufacturer of the insured automobile (but not the specific model). It is used together with another variable describing the specific model (not available in this dataset) to determine the type class for both motor TPL and motor vehicle own damage insurance. This variable therefore has no use for this study and will be ignored.

There is also information available about the year of first registration of the insured vehicle (ZULMMJJ1). Since FALTER represents the age of the vehicle similar

to how KUNALTER2 represents the age of the policy holder, ZULMMJJ1 will not be used for this study. For the variables VTYPH, NEUGE, KUEND and FELD62 detailed descriptions are not available. Thus, I will ignore them.

Altogether two new variables (KZFV and KUNALTER2) have been defined in this step and 14 variables have been excluded from the study. The remaining 28 variables (one response variable and 27 covariates) seem to be useful for the cancellation prophylaxis study.

Table A.1: Label, scale and range of the variables in the dataset

Variable	Label	Scale	Range
	Policy		
VSNR	policy number	nominal	0000001 - 9999999
VVNR	preliminary policy number	nominal	00000001 - 99999999
VERMNR	agent number	nominal	000001 - 999999
STORNO	cancelled?	nominal	0, 1
STODAT	date of cancellation	interval	ddmmyy, yy=96, 97, 98
STODAT1	year of cancellation	interval	96, 97, 98
BEGDAT1	year of first contract with VKB	interval	0 - 99
VWDAU	years in portfolio	ratio	0 - 99
	Policy holder		
GEBJJ	year of birth	interval	0 - 81
KUNALTER	age	ratio	18 - 99
KUNALTER2	age	ratio	18 - 99
AUSLK1	foreigner?	nominal	0, 1
PLZ	zip code	nominal	01000 - 99999
REGBE	county	nominal	1, 2, 3, 4, 5, 6, 7, 8, 9
ZART	payment option	ordinal	1, 2, 4
	Insured vehicle		
HERSTELL	manufacturer key	nominal	0001 - 9999
STAERKE	engine power (kW)	interval	1 - 999
ZULMMJJ1	year of first registration	interval	0 - 99
FALTER	age	ratio	0 - 99
KAUFDAT1	year of purchase	interval	0 - 99

(table continues)

Table A.1 (continued)

Variable	Label	Scale	Range
Motor TPL insurance			
KZKHT	tariff generation	nominal	O, 5, 6, A, B, C
DECKAR	type of coverage	nominal	3, 8, 9, 12, 52, 88
REGIO	regional class	nominal	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
TGR1	tariff group	nominal	A, B, C, D, K, N, R
KZWF	mileage per year	nominal	0, 1, 2, 3, 4, 5
FANZ1	number of drivers	nominal	0, F1, F2, J1, J2, N1, N2
KZEHB1	home owner?	nominal	0, 1
TYPKLH	type class	ordinal	0, 10 - 25
TARKLA	no-claims class	nominal	0 - 99
BEITSAT	percentage rate	ratio	30 - 275
JTBH	annual permium	ratio	0 - 9999
Motor vehicle own damage insurance			
KZFV	vehicle insurance?	nominal	0, 1
TYPKL	type class	ordinal	0, 10 - 40
FELD66	no-claims class	nominal	0 - 99
FELD68	percentage rate	ratio	0, 30 - 190
JTBF	annual premium	ratio	0 - 99999
Others			
SORAB1	special discount?	nominal	0, 1
VERBU1	employee tariff?	nominal	0, 1
VTYPH	preferred tariff	ordinal	0, 1, 2, 3
NEUGE	new customer	nominal	0, 1, 2
KUEND	cancellation in case of a claim	nominal	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
FELD62	characteristic second tariff	ordinal	000 - 999

APPENDIX B

RESULTS OF THE VARIABLE SCREENING

RESULTS OF THE VARIABLE SCREENING

This appendix contains some important results from the variable screening discussed in section 5.1.1. I will discuss the tariff generations in motor TPL in detail, and also highlight some of the results for both categorical and continuous covariates, including the definition of new variables.

B.1 Tariff Generations in Motor TPL insurance

The variable KZKHT represents the type of tariff used for the calculation of the premium for motor TPL insurance. The values O, 5 and 6 denote tariffs based on engine power used before 1996. A, B and C represent tariffs based on the type class of the insured vehicle used after June 1996; they were introduced on 6/1/1996, 1/1/1997 and 7/1/98 respectively.

For the value C, the proportion of cancelled contracts is very small. This is mostly due to the fact that this tariff generation was introduced in July 1998 and the sample of cancelled contracts is taken from the years 1996 - 1998; the not cancelled contracts, on the contrary, are all sampled in 1999. This is clearly a flaw of the sampling method and will limit the interpretability of the results.

Furthermore, the set of possible values for some other variables related to motor TPL depends on the tariff generation KZKHT. For example, the variable TYPKLH is equal to 0 for all motor TPL contracts for which the tariff is based on vehicle power (9,056 contracts in the sample); these contracts are those with values O, 5 or 6 for KZKHT.

Also the value for KZEHB1 (home owner discount) depends on the value for KZKHT since this type of discount is only offered in tariff generations A, B and C. More examples for these dependencies among predictor variables are given below (see variables DECKAR, KZWF and FANZ1).

Apart from that, the categories O, 5 and 6 may be combined since they all represent tariffs based on vehicle power and there is no further information about differences between these tariffs. This defines a new variable KZKHT2 with 4 levels. Some of the other categorical variables in the dataset are also worth closer examination.

B.2 Categorical Covariates

B.2.1 PLZ and PLZX (zip code areas)

The original variable PLZ (five digit zip code) has 2,668 different values in the sample dataset which means that it is not useful for the prediction of the cancellation behavior since all counts are small. However, the first two digits of the zip code represent larger areas in Germany, where big cities form separate areas. It may therefore be advisable to use the first two digits of the zip code (95 different values). For many of these areas the counts are very small, since the VKB only operates in Bavaria and in parts of Rheinland-Pfalz (22 zip code areas). I defined a new variable PLZX which represents the first two digits of the zip code for contracts in any of these 22 zip code areas; PLZX is set to zero for all other contracts.

B.2.2 DECKAR and DECKAR2 (type of coverage TPL)

The type of coverage for motor TPL insurance chosen by the policy holder is coded in DECKAR: 12 means unlimited coverage, 52 denotes unlimited coverage plus additional services, and the other values (3, 8, 9 and 88) represent different types of limited coverage ($< 1\%$ of the contracts). Hence I combined the values 3, 8, 9 and 88 and defined a new variable DECKAR2. The value 52 can only occur if KZKHT is equal to C. Overall, this variable seems to have limited usefulness.

B.2.3 KZWF, KZWF2 and KZWF3 (mileage per year)

KZWF represents the annual mileage restriction for the vehicle (see section 1.2.2): 0 means that no mileage restrictions are applicable for this contract, and it is equivalent with the mileage class 5. Therefore a new variable KZWF2 was defined by replacing 0 with 5. All contracts based on vehicle power (KZKHT equal to O, 5 or 6) have a value of 5 for KZWF2 since this discount was not offered in this tariff generation. Contracts in tariff generations A or B can either have a discount (KZWF2 = 1) or not (KZWF2 = 5). The values 2, 3 and 4 can only occur if KZKHT is equal to C and represent different levels of discount. Further investigation has shown that it seems advisable to define a new variable KZWF3 which indicates whether there is a discount for restricted mileage (KZWF2 = 1, 2, 3 or 4) or not (KZWF2 = 5).

B.2.4 FANZ1, FANZ3 and FANZ4 (number of drivers)

If the number of drivers insured under the contract is limited, a discount on the basic premium may apply (see section 1.2.2). These restrictions are stored in the variable FANZ1, where 0 means no limitations, N1, J1 and F1 denote restriction on one driver and

N2, J2 and F2 denote restriction on two drivers. Discounts apply for the values F1/F2 and J1/J2, but not for N1/N2. The 7×2 crossclassification table of FANZ1 versus the outcome STORNO has zero counts for F1 and F2; these values can only occur when the contract is in tariff generation C (variable KZKHT). Since J1/J2 and F1/F2 apply in different tariff generations but have essentially the same meaning (discount for one/two drivers), the variable FANZ4 was defined by setting all values F1 (F2) equal to J1 (J2). In addition, a variable representing whether there is a restriction on the number of drivers (FANZ4 $\neq 0$) or not has been found useful as a predictor (binary variable FANZ3).

B.2.5 TGR1 and TGR2 (tariff group motor TPL)

For the variable TGR1, the contingency table of TGR1 versus the binary outcome STORNO has zero counts for the values C, D, K and R; they represent professional groups other than the main groups A, B and N (see section 1.2.2). Since they together only represent less than 2% of the observations, we combine them into one category O (other professional groups) and define a new variable TGR2.

B.3 Continuous Covariates

An overview of the continuous variables is given in Table B.2. For all variables related to vehicle insurance (TYPKL, TARKLAFV, FELD68 and JTBF), a value of zero means that vehicle insurance is not included in the contract. For TYPKLH (type class motor TPL) a value of zero means that the premium calculation for motor TPL is based on an old tariff for which vehicle power is used as a tariff factor instead of type class.

For the computation of the quantiles for all these variables, observations with a value of zero have been excluded.

Positive values for JTBF represent the annual premium for partial or full vehicle insurance, depending on the choice of the customer. Premiums for the partial coverage tend to be low, whereas premiums for the full coverage can be very high. Customers usually have full coverage only as long as the car is relatively new. Hence the premium level may be related to the age of the vehicle. However, there is no information provided about the type of coverage (partial or full) chosen.

Table B.1: Range and quantiles for continuous variables

Variable	Range	Quantiles				
		1%	5%	50%	95%	99%
VWDAU	0 - 41	0	0	6	25	28
KUNALTER2	18 - 97	22	27	45	73	82
ZART	1, 2, 4	1	1	2	4	4
STAERKE	7 - 280	25	33	57	113	150
FALTER	0 - 63	0	1	8	17	22
TYPKLH	0, 10 - 25	11	12	16	21	23
TARKLA2	0, 1 - 34	1	1	11	29	34
BEITSAT	30 - 275	30	30	45	120	240
JTBH	25.0 - 3155.3	491.3	624.6	1123.8	1770.2	2153.7
TYPKL	0, 10 - 40	11	13	19	33	37
TARKLAFV	0, 1 - 34	1	1	1	26	33
FELD68	0, 30 - 190	30	30	100	100	100
JTBF	0, 26.4 - 12125.0	45.2	70.2	273.85	2028.2	3065.1

APPENDIX C

LOGISTIC REGRESSION MODELS IN SAS

LOGISTIC REGRESSION MODELS IN SAS

This appendix contains the SAS code used for the assessment of logistic regression models via external validation.

C.1 Fitting of the Model

```
TITLE 'Model assessment:  22 variables';

* Choose partition!;
* Y = STORNO for the training sample and
* Y is missing for the validation sample;
DATA work.total4;
    SET work.total4;
    Y = Y1;
KEEP _ALL_;
RUN;

* Fit logistic regression model!;
* Model probability of cancellation;
PROC LOGISTIC DATA=work.total4 DESCENDING;
   CLASS FALTER0 (REF='No')        KZKHT2 (REF='01/97 - 06/98')
         REGIO (REF='0')           TGR2 (REF='Normal')
         KZWF3 (REF='No')          BEITSAT2 (REF='No')
         FANZ3 (REF='No')          KZEHB1 (REF='No')
         KZFV (REF='Yes')          VERBU1 (REF='No')
         / PARAM=REF;

    MODEL Y = VWDAU STAERKE FALTER FALTER0 KZKHT2 TYPKLH REGIO TGR2
              KZWF3 BEITSAT3 BEITSAT2 FANZ3 KZEHB1 JTBH KZFV JTBF
              VERBU1 KZKHT2*STAERKE KZKHT2*JTBH KZKHT2*TYPKLH
              STAERKE*JTBH TYPKLH*JTBH;

    OUTPUT OUT=work.pred XBETA=logit PROB=phat;
    * Output dataset contains the same variables as the input data set;
    * Additionally, it contains the predicted logits and;
    * the predicted probabilities;
RUN;
```

C.2 Pearson Chi-Square Test

```
TITLE 'Normal approximation to the Pearson Chi-Square Statistic';

* Consider validation sample!;
* Calculate squared Pearson residual and variance term!;
DATA work.pearson1;
    SET work.pred;
    WHERE Y = .  and phat ne .;

    r = (STORNO-phat)**2 / (phat*(1-phat));
    v = (1/(phat*(1-phat))) - 4;

    KEEP _ALL_;
RUN;

* Find Pearson Chi-Square Statistic and variance sigma_v**2!;
PROC UNIVARIATE DATA=work.pearson1 NOPRINT;
    VAR r v;
    OUTPUT OUT=work.pearson2 SUM=Pearson Sigma_2 N=n_v;
RUN;

* Calculate Z statistic and find two-tailed p-value!;
DATA work.pearson2;
    SET work.pearson2;

    Z = (Pearson - n_v) / sqrt(Sigma_2);
    p = 2 * ( 1 - cdf('normal',abs(Z)));

    LABEL Pearson='Pearson chi-square statistic'
          Sigma_2='Variance (sigma_v**2)' n_v='# obs (n_v)'
          Z='Standardized Statistic Z' p='p-value (two-tailed)';
    KEEP _ALL_;
RUN;

* Print results!;
PROC PRINT DATA=work.pearson2 LABEL NOOBS;
    VAR n_v Pearson Sigma_2 Z p;
RUN;

* Clean up!;
PROC DATASETS LIBRARY=work;
    DELETE pearson1 pearson2;
RUN;
```

C.3 Stukel's Test

```
TITLE 'Stukels Test';

* Compute new variables z1 and z2 from fitted values!;
DATA work.pred;
    SET work.pred;

    IF phat >= 0.5 THEN DO;
        z1 = 0.5 * logit**2;
        z2 = 0;
        END;
    ELSE DO;
        z1 = 0;
        z2 = - 0.5 * logit**2;
    END;
    KEEP _ALL_;
RUN;

* Include all 22 variables!;
* Test z1 and z2 for addition to the model!;
* Look at Residual chi-square test (score test)!;
PROC LOGISTIC DATA=work.pred DESCENDING;
   CLASS FALTER0 (REF='No')        KZKHT2 (REF='01/97 - 06/98')
         REGIO (REF='0')           TGR2 (REF='Normal')
         KZWF3 (REF='No')          BEITSAT2 (REF='No')
         FANZ3 (REF='No')          KZEHB1 (REF='No')
         KZFV (REF='Yes')          VERBU1 (REF='No')
         / PARAM=REF;

    MODEL Y = VWDAU STAERKE FALTER FALTER0 KZKHT2 TYPKLH REGIO TGR2
              KZWF3 BEITSAT3 BEITSAT2 FANZ3 KZEHB1 JTBH KZFV JTBF
              VERBU1 KZKHT2*STAERKE KZKHT2*JTBH KZKHT2*TYPKLH
              STAERKE*JTBH TYPKLH*JTBH
              Z1 Z2
              / SELECTION=FORWARD SLENTRY=0 INCLUDE=22 DETAILS;
RUN;
```

C.4 Hosmer-Lemeshow Test

```
TITLE 'Hosmer-Lemeshow Test (validation sample):  deciles of risk';

* Define grouping variable!;
DATA work.hltest;
    SET work.pred;
    WHERE Y = .  and phat ne .;
    phatgroup = phat;
    KEEP STORNO phat phatgroup;
RUN;

* Divide into 10 groups of equal size according to predicted
* probabilities and sort!;
PROC RANK DATA=work.hltest GROUPS=10 OUT=work.hltest;
    VAR phatgroup;
RUN;

PROC SORT DATA=work.hltest;
    BY phatgroup;
RUN;

* Find total # of observations and sum of observations per group
* for phat and response STORNO!;
PROC UNIVARIATE DATA=work.hltest NOPRINT;
    BY phatgroup;
    VAR STORNO phat;
    OUTPUT OUT=work.hltest2 N=n_y n_p SUM=sum_y sum_p;
RUN;

* Calculate components of H-L statistic!;
DATA work.hltest2;
    SET work.hltest2;

    group = phatgroup + 1;
    num = (sum_y - sum_p)**2;
    denom = sum_p * (1 - sum_p / n_p);
    quotient = num / denom;
    pvalue = 1 - cdf('chisquare',quotient,1);

    LABEL quotient='C_k' group='Group' sum_y='Observed'
          sum_p='Expected' n_y='Total' pvalue='p-value (df=1)';
    KEEP _ALL_;
RUN;
```

```
* Calculate Hosmer-Lemeshow Test Statistic!;
PROC UNIVARIATE DATA=work.hltest2 NOPRINT;
    VAR quotient;
    OUTPUT OUT=work.hltest3 SUM=sum;
RUN;

* Find p-value from Chi-Square distribution!;
DATA work.hltest3;
    SET work.hltest3;

    p = 1 - cdf('chisquare',sum,10);

    LABEL sum='Hosmer-Lemeshow Statistic' p='p-value (df=10)';
    KEEP _ALL_;
RUN;

* Print results!;
PROC PRINT DATA=work.hltest2 NOOBS LABEL;
    VAR group n_y sum_y sum_p quotient pvalue;
    SUM n_y sum_y sum_p quotient;
RUN;

PROC PRINT DATA=work.hltest3 NOOBS LABEL;
RUN;

* Clean up!;
PROC DATASETS LIBRARY=work;
    DELETE hltest hltest2 hltest3;
RUN;
```

C.5 Classification Tables

```
TITLE 'Classification table (validation sample)';

* Use macro to compute classification table for several cutpoints;
%LET response = STORNO;
%LET pred = ypred;
%LET respyes = 1;
%LET respno = 0;

%MACRO ctable(num,step);
* num=number of cutpoints;
* step=step size between cutpoints;
```

```
* Classify observations in the validation sample using all
* cutpoints in the sequence!;
DATA work.validate;
    SET work.pred;
    WHERE Y = .  ;
    ARRAY pt(&num) 8 &pred.1-&pred.&num;

    DO j=1 to &num;
        IF phat = .  THEN pt(j) = .;
        ELSE IF phat >= j*&step THEN pt(j) = &respyes;
        ELSE pt(j) = &respno;
    END;

    KEEP &pred.1-&pred.&num &response phat;
RUN;

%DO n=1 %TO &num; * for every single cutpoint ...;

    * Compute classification table (predicted vs.  observed)!;
    PROC FREQ DATA=work.validate NOPRINT;
        TABLES &pred.&n*&response / OUT=work.freqs OUTPCT;
    RUN;

    %LET yy = 0; /* predicted YES, observed YES */
    %LET yn = 0; /* predicted YES, observed NO */
    %LET ny = 0; /* predicted NO, observed YES */
    %LET nn = 1; /* predicted NO, observed NO */
    %LET sens = 0; /* sensitivity */
    %LET spec = 0; /* specificity */
    %LET fpos = 0; /* false positive rate */
    %LET fneg = 0; /* false negative rate */

    * Sort classification table (order:  yy, yn, ny, nn)!;
    PROC SORT DATA=work.freqs OUT=work.freqs;
        BY DESCENDING &pred.&n DESCENDING &response;
    RUN;

    * Get information from classification table!;
    DATA work.freqstemp();
        SET work.freqs;
        WHERE &pred.&n NE .;
        LENGTH mcr sens spec fpos fneg yy yn nn ny cprob 8;
```

```
    IF &pred.&n = &respyes AND &response = &respyes THEN
    DO;
        CALL symput('yy',COUNT) ; /* save in macro variable */
        CALL symput('sens',PCT_COL);
    END;
    ELSE IF &pred.&n = &respyes AND &response = &respno THEN
    DO;
        CALL symput('yn',COUNT);
        CALL symput('fpos',PCT_ROW);
    END;
    ELSE IF &pred.&n = &respno AND &response = &respyes THEN
    DO;
        CALL symput('ny',COUNT);
        CALL symput('fneg',PCT_ROW);
    END;
    ELSE IF &pred.&n = &respno AND &response = &respno THEN
    DO;
        CALL symput('nn',COUNT);
        CALL symput('spec',PCT_COL);
        yn=symget('yn'); /* get value from macro variable */
        ny=symget('ny');
        nn=symget('nn');
        yy=symget('yy');
        mcr = (yn+ny)/(yn+ny+yy+nn);
        spec=symget('spec');
        sens=symget('sens');
        fpos=symget('fpos');
        fneg=symget('fneg');
        cprob=symget('n');
        cprob=cprob*&step;
    END;

    KEEP mcr sens spec fpos fneg cprob;
RUN;


%IF &n = 1 %THEN /* first cutpoint */
%DO;
    DATA work.freqall;
        SET work.freqstemp;
        WHERE mcr NE .;
    RUN;
%END;
```

```
    %ELSE %DO;
        DATA work.freqstemp;
            SET work.freqstemp;
            WHERE mcr ne .;
        RUN;

        PROC APPEND BASE=work.freqall DATA=work.freqstemp;
        RUN;
    %END;
%END;

PROC DATASETS LIBRARY=work;
    DELETE freqs freqstemp validate;
RUN;

%MEND ctable;

* Run macro!;
%ctable(19,0.05);

* Set up classification table!;
DATA work.ctable;
    SET work.freqall;
    LABEL sens='Sensitivity' spec='Specificity'
          fpos='False positive rate' fneg='False negative rate'
          mcr='Misclassification rate' cprob='Cutpoint'
          _1mspec='1 - Specificity';

    _1mspec = 100 - spec;

    KEEP _ALL_;
RUN;

* Print classification table!;
PROC PRINT DATA=work.ctable LABEL NOOBS;
    VAR cprob mcr sens spec fpos fneg ;
RUN;

* Clean up!;
PROC DATASETS LIBRARY=work;
    DELETE freqall ctable;
RUN;
```

C.6 Area under the ROC Curve

```
TITLE 'Area under the ROC curve (validation sample)';

* Consider validation sample!;
DATA work.wilcoxon;
    SET work.pred;
    WHERE Y = .  and phat ne .;
    KEEP STORNO phat;
RUN;

* Find Wilcoxon rank sum statistic _WIL_!;
* Two groups are defined by binary outcome STORNO!;
* Computes rank sum for smaller group!;
PROC NPAR1WAY DATA=work.wilcoxon WILCOXON NOPRINT;
    CLASS STORNO;
    VAR phat;
    OUTPUT OUT=work.wilcoxon2 WILCOXON;
RUN;

* Find number of positive reponses and total number of
* observations in validation sample!;
PROC SORT DATA=work.wilcoxon;
    BY DESCENDING STORNO;
RUN;

PROC UNIVARIATE DATA=work.wilcoxon NOPRINT;
    VAR STORNO;
    OUTPUT OUT=work.wilcoxon3 SUM=sumyes N=total;
RUN;

* First calculate Mann-Whitney U statistic by;
* subtracting m*(m+1), where m=# of obs in smaller group!
* Then divide by number of pairs!;
DATA work.wilcoxon4;
    SET work.wilcoxon3;
    SET work.wilcoxon2;

    sumno = total - sumyes;
    IF sumyes < sumno THEN DO;
        U = _WIL_ - sumyes*(sumyes+1)/2;
        ROC = U / (sumyes*sumno);
    END;
```

```
    ELSE DO;
        U = _WIL_ - sumno*(sumno+1)/2;
        ROC = 1 - U / (sumyes*sumno);
    END;

    LABEL ROC='Area under the ROC curve';
    KEEP total sumyes sumno _WIL_ U ROC;
RUN;

* Print result!;
PROC PRINT DATA=work.wilcoxon4 NOOBS LABEL;
    VAR ROC;
RUN;

* Clean up!;
PROC DATASETS LIBRARY=work;
    DELETE wilcoxon wilcoxon2 wilcoxon3 wilcoxon4;
RUN;
```

APPENDIX D

CLASSIFICATION TREES IN SAS

CLASSIFICATION TREES IN SAS

In this appendix I will describe how the SAS Enterprise Miner can be used to build classification trees as discussed in chapter 4. For details see the SAS help documentation.

D.1 Selection of Splits

By setting the "maximum number of branches from a node" equal to 2, the set of possible splits is restricted to binary splits on a single variable. Linear combination splits are not available in this implementation. Both Gini index and entropy measure of impurity are available for the selection of the splits. Since the dataset is not too large, all observations should be used for the split search (no subsampling). Therefore the number of "observations necessary for a split search" is set to 13,333, the number of observations in the training set. For binary splits on binary targets, the optimal split is always found and we can set the "maximum tries in an exhaustive split search" such that all possible splits are enumerated (for example, 99,999,999). Several (for example, 5) competing "splitting rules are saved in each node" and may be used for comparison of competing splits and interpretation.

D.2 Pruning

As discussed in section 4.3, first a large tree is grown. By setting the "minimum number of observations in a leaf" equal to 1, we make sure that pure nodes are not split. To grow the largest tree first, we set the "number of observations required for a split search" equal to $N_{min} + 1$ (for example, equal to 2). Since the covariate vectors $\mathbf{x_i}$ in

the dataset are all distinct, it is impossible that a node with more than one observation contains only identical covariate vectors. There is also a restriction on the "maximum depth of a tree" (100 levels), but this will not be a problem in this study.

The SAS Enterprise Miner uses the first algorithm discussed in section 4.3 to prune the large tree, based on the misclassification rate on the training sample. It is not possible to use cost-complexity pruning. The smallest subtree with the smallest misclassification rate on the validation sample is then chosen. In order to use this procedure the "model assessment measure" has to be set to the proportion correctly classified and the "subtree method" to the best assessment value. The "One standard error rule" has to be implemented manually. Crossvalidation is not available.

D.3 Priors and Misclassification Costs

Priors and misclassification costs can be defined in the "target profile". There are two options which allow the use of priors and/or misclassification costs during the split search.

D.4 Surrogate Splits

Surrogate rules as described in section 4.5 can be computed (by setting the option "surrogate rules saved in each node" equal to the number of variables minus 1). Missing values are not used as separate categories in the tree construction if the checkbox "treat missing as an acceptable value" is disabled. The importance measure computed by the SAS Enterprise Miner is slightly different from the one originally suggested by Breiman et al. (1984) and was discussed at the end of chapter 4.

D.5 Output

The output generated for tree models includes a graphical display of the "best" subtree chosen and the corresponding classification table for both training and validation sample. Also the node definition, competing splits and surrogate splits for each node can be examined. Another graphical display is a tree ring, which illustrates tree complexity, split balance, and discriminatory power. The center region of the ring represents the entire data set (the root node of the tree). The ring surronding the center represents the initial split. Successive rings represent successive stages of splits, where the sizes of displayed segments in one ring are proportional to the number of training observations in the segments. The last ring corresponds to the final partition of the covariate space. There are several options for the coloring of a tree ring.

ABSTRACT

ABSTRACT

The deregulation of the German insurance industry in 1994 has significantly intensified competition between insurance companies, especially in the important sector of motor insurance. One consequence is that insurance companies are interested in identifying contracts in their portfolio that are likely to be cancelled. This would enable them to take appropriate measures with the goal of preventing these customers from cancelling (cancellation prophylaxis).

For this purpose a cancellation prophylaxis study was designed, and a random sample of 20,000 contracts from the portfolio of a large German insurance company was taken. In this thesis, two modeling approaches for the prediction of the cancellation behavior are considered: logistic regression models and classification trees based on the CART methodology.

For both methods, the basic methodology and some recent developments are summarized. The model selection and model assessment processes are described, and main results are discussed. We conclude with a comparison of logistic regression models and classification trees, as well as suggestions for improvement of the study. Appendices contain detailed information on the provided dataset and on the implementation of the methods in SAS.