POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Matematica

Tesi di laurea Magistrale

Predictive methods for calculating the non-life insurance premium

Relatore

prof. Francesco Vaccarino

Corelatore

prof. Paolo Brandimarte

Giulia Rocca

matricola 244144

Anno accademico 2018-2019

Tesi: Predictive methods for calculating the non-life insurance premium

Candidato: Giulia Rocca Matricola: s244144

Relatore: Francesco Vaccarino Corelatore: Paolo Brandimarte

Contents

Introduction

In order to protect against the risks inherent in an economic or professional activity, or against the risks in which it is possible to incur in private life, businesses and families have the possibility to sign an insurance policy. The latter is a contract with which the insurer undertakes to compensate the contractor for damages suffered as a result of an accident, during the period of coverage and according to the procedures established by the contract. Obviously, the transfer of risk is carried out only against payment of an appropriate compensation: the premium.

The determination of the premium to which insurance companies offer the policies is one of the most important phases of the insurance process, because the correct definition of premium allows the companies to meet the commitments taken with its policyholders and to guarantee an adequate remuneration of capital.

The process of defining the premium for a non-life insurance policy, starts from the probabilistic assessment of the total amount of damage expected, caused by the claims during the insured period. This premium configuration is called *fair premium* and is the starting point on which to build the rate. At fair premium, in fact, the company must add a profit for the service offered during the period of coverage, thus obtaining the *pure premium*. Finally, fixed charges and taxes are added to get the final price of the insurance policy,that is the *gross premium*.

In this thesis, the objective is to illustrate the procedure for determining the fair premium, using Generalized Linear Models, and the pure premium in a non-life insurance, through the optimization of appropriate functions.

Now follows a short description of how it was decided to articulate this thesis work.

In Chapter 1 the notion of insurance was introduced and the issue of non-life insurance discussed, focusing attention on characteristics of the Car Liability insurance. The non-life pricing process was then presented, starting with the notion of insurance premium. The three typical premium configurations have been defined, namely the *fair premium*, the *pure premium* and the *tariff premium*. At the end of chapter, some information was also provided on the tariff classes and the personalization of premium.

Chapter 2 is dedicated to Generalized Linear Models, widely used in actuarial practice for the construction of non-life premiums. After briefly presenting the structure of these models, the models for the *number of claims*, with Poisson distribution, and the models for *damage per claim*, with Gamma distribution, were described.

In Chapter 3 an empirical application was made, starting from data provided by RGI S.p.A. The chapter has been divided into three parts: in the first part, useful Machine Learning techniques were analyzed, in the second part a presentation of the data was made and finally, in the third part, the model for the definition of fair premium was developed using GLMs.

In Chapter 4, the pricing strategies that can be adopted by insurance companies to find the optimal gain are described. In particular, two different strategies have been considered: the first, which takes into account only the customers and the demand linked to them, and the second, which also considers the competitors and the reputation of the company. Finally, it has been described how the pricing of insurance changed with the advent of the Internet.

Chapter 1

Non-life Insurance: Products and Pricing

1.1 Insurance products

A short description of the main features of insurance, in particular of non-life insurance, is provided in this Section, mainly aiming at introducing the basic items involved in premium calculation.

1.1.1 General Aspects

According to the Art. 1882 of Civil Code, «insurance is the contract with which an insurer, in exchange of the payment of a certain premium, obliged himself to pay an indemnity to the insured equivalent to the damage caused by an accident and to pay an income or a capital if a life-related event occurs».

The article therefore introduces the distinction between *non-life* insurance (also named general or property/casualty insurance) and *life* insurance. With regard to non-life insurance, there is an obligation for the insurer to indemnify the insured from the damages suffered, due to an unfavorable event; while in the case of life insurance, the law states that the insurer must provide for the payment of a capital or an annuity, if an event related to human life takes place. In this thesis, only non-life insurance is considered.

The law provides that two or more parties agree through a contract, represented by a policy, to build and regulate a specific legal relationship patrimonial,

that is the insurance. Function of the contract coincides with the elimination of risk and this purpose can be achieved by trasferring the risk to insurer, i.e. a specialized entity which carries out the insurance business in an entrepreneurial manner. Insurer is the one able to neutralize the risk that has taken, with the signing of single contract, through its inclusion in a group of risks (*pool* of risks). Pool contains risks that have homogeneity, size and independence as their essential characteristic and thanks to these peculiarities, the company is able to transform what is uncertain for the individual insured, in certainty for the mass of risks managed. Then the object of policy is the risk: contract is void if the risk does not exist or has ceased before the conclusion of the contract (art. 1895 of the Civil Code). To allow the insurance company to make a correct assessment of the risk, it is of primary importance that the insured provides in a precise manner all the necessary data.

On one side, in a contract there is therefore a subject who professionally accepts to assume an economic risk upon payment of a prize, committing to perform the service at the moment in which the feared event will occur. On the other, there is the insured, who represents the person who is potentially subject to an unfavorable event,but that, thanks to the stipulation of insurance contract, protects himself from the risk in question, by paying a price.

At the signing of contract, the parties can not know if the insured risk will occur during the warranty period. For this reason, the insurance contract falls into the category of *random* contracts: on the one hand, the insured person is not able to know whether he will receive indemnification or capital from the insurer, against the premium paid in advance; on the other hand, the insurer ignores whether he will have to pay the promised benefit or if the premiums collected against the risks assumed will be adequate to meet the payments due.

Thanks to the progressive improvement of the statistical data, today the insurer is able to share, with ever greater precision, the risks among insured. However, even if the insurer's estimates were entirely correct, the random nature of the contract would persist. In fact, the insurance company would know the exact number of claims that will affect the mass of insured risks and the exact cost of the services it will have to pay, but in any case it would not be able to know for which contracts it will be obliged subsequently to perform the service. Random

character is therefore a peculiar element of the insurance contract (Miani, 2010) and, due to uncertainty of performances, non-life contracts generally have a short coverage period, typically one year, called *policy year*.

1.1.2 Main Categories of Non-life Insurance Products

Non-life insurance includes a wide range of products, offering protection in respect of many risks. This thesis is not intended to provide a complete and detailed presentation of the possible contents of non-life insurance coverage; some informations which are useful to understand the fundamentals of pricing will be given.

The non-life business may be segmented according to different perspectives. Considering the possible contractor, it can be distinguished in *personal* insurance, addressed to individuals or families (e.g. motor insurance, health insurance, homeowners insurance, and so on), and *commercial* insurance, useful to business entities (e.g. transportation insurance, workers compensation, and so on). In relation to the possible beneficiary, insurance can be also classified into: *property* insurance, *liability* (or casualty) insurance and *health* insurance.

Property insurance provides a protection against a possible loss or damage to the property of the insured, including loss of profits or emergence of costs. Insurances such as Fire, Theft, Transport, Hail, Judicial Protection and Boats are part of property insurances. For these insurances, an amount called *insured capital* is determined, aimed at dealing with damaging events. This sum coincides with the maximum indemnity paid by the insurer in the event of a claim and is usually commensurate with the value of asset.

Health insurance offers instead the payment of a compensation by insurer, in the event of an accident or illness of the insured, or indemnity, in the event of death. *Accident* policies can be divided, depending on the type of accident, into: temporary disability, permanent disability and death. In the case of temporary disability the insurer will pay a daily allowance whose amount is fixed; in the case of permanent disability, the benefit payable by the insurer is calculated as a percentage of the sum insured on the basis of percentage of disability caused by the claim; in the case of death, will be paid a fixed sum that was set at the time the contract was signed. The benefits provided for the *Sickness* insurance policies are

instead reimbursements of medical expenses, paid by the insurer within a certain ceiling. It is noteworthy that some forms of health insurance, typically those with forfeiture benefits and a duration of more than one year, are classified within life insurance.

Liability insurances offer financial protection against various damage caused by the insured to third parties. The most important of these, in economic terms, is Car Liability (R.C.A in Italian). For liability insurance a maximum amount is defined, i.e. the limit within which the insurer intervenes to compensate the damage caused by the insured party to third parties. The maximum amount limit has the objective to contain, within a reasonable threshold, the possible risk exposure of the insurer.

While the general principles for pricing and reserving are common to all the business lines, the specific methods applied in practice may differ significantly, consistent with the features of particular line of business dealt with.

Car Liability in Italy

As mentioned above, Car Liability insurance is the most important branch in the non-life insurance sector. It is a compulsory liability insurance and given its peculiarities and its economic importance, it forms a separate branch within the non-life insurance.

Since 1969, anyone who puts a vehicle in Italy in circulation has the obligation to take out an insurance policy. The Car Liability in fact is the contract that has the purpose of guaranteeing the driver or, if different, the owner of vehicle, against the risk of having to compensate third parties for damages caused by the circulation of vehicle. This policy therefore covers the damage caused by the insured vehicle to people, animals or property as a result of a claim, but does not guarantee coverage for any physical damage suffered by the driver who caused the accident.

The duration of contract is one year, starting from midnight on the day prize has been paid. It is possible to issue policies with a duration of less than one year (called *temporary*), in particular for vehicles with temporary license plates and those calculating for testing or demonstration. The company is obliged to

compensate claims incurred by the due date of the policy.

Another characteristic of a Car Liability was the presence of tacit renewal, i.e. the automatic renewal of coverage for a further annuity in the absence of withdrawal by the insured, but this clause has been abolished with the law decree number 179. Moreover, a "tolerance" period of 15 days beyond the expiration date has been introduced, during which the company continues to respond to claims caused by the insured. Tolerance period allows the insured to evaluate the different offers on the market and decide whether to keep the same company or change it. This decree facilitated the review of the prices proposed on the market thanks to the greater competition among the insurance companies.

One of elements that influence the premium of these policies is the *Bonus Malus* mechanism, introduced with the Bersani law. It provides the adjustment of premium on the basis of individual experience, depending on the number of claims caused during the period of validation of the contract and on basis of a parameter linked to a *merit class*, which measures the claims of previous years.

Each company may choose to adopt specific rules in the allocation of internal merit class to its customers, without prejudice to the obligation to provide for the correspondence rules between the internal merit classes and the universal merit classes. The companies have the obligation to report not only the internal merit class, but also the universal merit class in the risk certificate.

The universal merit classes are 18: the 18th is the highest, the one at which the highest prize will be paid; 1st is the lowest class, where the prize is cheaper. Therefore the lower the class, the cheaper the policy. The 14th class is generally assigned to vehicles insured for the first time after matriculation or transfer of ownership.

The merit class improves (*bonus*) if no accidents occur within the insured year, instead it worsens (*malus*) in the presence of claims for which the driver is responsible and that have been paid by the insurer during an *observation period*. The bonus case provides for the improvement of a merit class, benefiting from the reduction of premium for lowering the class, while in the case of malus two classes are increased, with a consequent increase in premiums. The observation period is the period of time in which insurer evaluates the driver's driving behavior. It starts from the day the policy begins, and then finish 60 days before the annual

deadline[1].

As mentioned before, the merit classes adopted by various companies are not all the same, since they are left with a margin of autonomy in the management and definition of internal classes. In order to ensure however comparability between the various systems adopted by the companies, IVASS, that is an insurance supervision institute, introduced the universal conversion merit class (CU).

Figure 1.1 shows the possible class displacements compared to the departure one, in the case of one or more claims reported during the policy observation period.

Starting CU	0 claims	1 claim	2 claims	3 claims	4 claims or more
1	1	3	6	9	12
2	1	4	7	10	13
3	2	5	8	11	14
4	3	6	9	12	15
5	4	7	10	13	16
6	5	8	11	14	17
7	6	9	12	15	18
8	7	10	13	16	18
9	8	11	14	17	18
10	9	12	15	18	18
11	10	13	16	18	18
12	11	14	17	18	18
13	12	15	18	18	18
14	13	16	18	18	18
15	14	17	18	18	18
16	15	18	18	18	18
17	16	18	18	18	18
18	17	18	18	18	18

Figure 1.1: Table of Universal Classes based on the number of claims occuring in a year

In the insurance sector, the Bersani Decree allows the owner of a new or used vehicle to acquire the same class of merit of a circulating and insured vehicle already in his possession or in possession of a member of the cohabiting family unit. Thanks to this mechanism, even new drivers who make insurance for the

[1]For coverage after the first year, the observation period begins two months before the start of the contract and ends two months before the deadline.

first time can avoid starting from the fourteenth class of merit (or CU), inheriting that of the parent.

The universal merit class affects the definition of premium to be paid, but using the same category does not mean paying the same insurance premium. The price of the RCA, in fact, also depends on other factors, including the personal characteristics of the figures listed in the policy, the number of years from which it is licensed, the power and fuel type of vehicle.

If the car has a regular insurance coverage, any damage caused by a claim will be reimbursed by the company within the limits set by a maximum amount (agreed upon when the policy contract was signed) that the insurance compensates in the event of an accident.

Starting from 11 June 2017, following the adaptation of the minimum legal limits established by European legislation 2009/103/CE and the Private Insurance Code (art. 128), higher maximum amounts have been set to guarantee insured drivers a greater coverage in the event of damage. The law provides for 7.290.000€ as a single minimum maximum amount; if the contractor decides to insert the two separate maximum amounts, the minimum statutory maximum amount for damage to property is 1.220.000€, while that for personal injuries is set at 6.070.000€. Beyond these thresholds the insured will respond with his own assets. All companies can also offer higher limits by charging the person who stipulates a higher insurance premium.

Finally, a further peculiar element, which concerns Car Liability, is the method of claim compensation. There are two different compensation procedures: the *ordinary* procedure and the *direct* compensation procedure.

The direct compensation procedure was introduced by the Private Insurance Code in February 2007. This procedure requires that the driver not responsible, or partially responsible for the claim, advances a claim for compensation to his insurance company. The procedure of direct compensation, compared to the ordinary one, allows to considerably reduce the settlement time of claims and this has a positive effect on the relationship between the insured and the company. It also allows a reduction in premiums over the medium/long term, as management and administration costs are reduced for insurance companies.

The direct compensation procedure applies basically to all road accidents

between two vehicles, except for those involving:

- more than two vehicles (think about the chain crashes);
- a vehicle not regularly insured or not registered in Italy (in this case the claim for compensation must be presented to the Guarantee Fund for victims of the road and to the company designated according to the place of occurrence of the accident);
- a vehicle that does not belong to the category of motor vehicles;
- a pedestrian, a cyclist or a real estate (think of the hypothesis of the driver who, losing control, goes to smash the window of a shop);
- a special vehicle or an agricultural machine.

Furthermore, the direct compensation procedure can not be activated when:

- there was no impact between the two vehicles;
- there are serious injuries leading to a permanent invalidity of more than 9% resulting from the accident.

In cases where the direct compensation procedure is not applicable, the ordinary compensation procedure will be followed. In this case, the claim for compensation must be made to the vehicle insurance company responsible for the accident.

1.2 Pricing in non-life insurance

The insurance company carries out a particular productive activity, represented by the systematic assumption of risks through insurance contracts. The contract establishes that the insured obtains a promise of compensation subject to the occurrence of an event, indicated in the policy. Obviously, the company will take on the risks of insured persons on payment of a compensation, i.e. the *insurance premium*.

From what has just been said, it can therefore be said that the premium is, by its nature, the proceeds of the insurance business carried out by the company,

while the reimbursements are instead the costs connected to the exercise of this activity, to which must also be added management and organization costs, as well as administrative burdens.

The sale of insurance product takes place before its production, so there is an inversion of the production cycle, induced by the precedence of revenues compared to the costs. In fact, premiums, which are collected in advance and represent revenues, accrue before compensation, which is instead the typical cost item. Contrary to what happens for a normal company, which produces goods or services in which the selling price of finished products is calculated on the basis of costs already incurred, the insurer is not able to define precisely when it offers its own market services and costs for performance, future and uncertain. Estimates must therefore be made.

Hence, the efficiency of management depends on the quality of probabilistic estimates and their applicability to the risk group or to the insured parties. The correct estimate of probabilities, from which the revenue capacity derives to cover costs, depends on the correct functioning and right application of the law of large numbers: the wider the homogeneous and independent risk sample observed, the more the estimated frequencies will tend to coincide with the probabilities of events. The goal for company is to be able to apply a "today" premium able to cover any "tomorrow" compensation. The collected revenues must also be able to guarantee an adequate remuneration of the company's capital, therefore it is not sufficient to cover the costs alone.

The inversion of the production cycle, as well as influencing the income statement, is also reflected in the company's capital structure.

The sale of policy, which provides for a future payment obligation, raises a contingent-state debt to insured. The onset of this debt requires appropriate investments in assets to support the values of these liabilities. The premium that company collects in advance must therefore be invested in order to guarantee the economic-patrimonial balance of the company. The objective is the achievement of a portfolio composition of assets that present an adequate combination of risk, yield and liquidity, able to cope with characteristics of the liabilities and profit objectives set by the company (Porzio, et al., 2011).

The level of "essentiality" of the investment varies according to the branch

of activity for which the company has received the authorization. In fact, for life insurance coverage, which has a multi-year duration, it is essential that the premiums collected are invested. In this case, therefore, the investment is not a simple economic opportunity, but a real necessity; the insurer who does not invest the premiums would risk ending up very quickly in a situation of insolvency or inability of the resources available to meet the third party's credit reasons. In the non-life classes, however, the brevity of coverage, which generally lasts one year or less, does not make it necessary for the company to invest premiums collected in order to achieve economic and financial equilibrium. However, the policy of mere custody of premiums would soon prove to be detrimental due to the deviations that can be ascertained between estimated frequencies and frequencies detected. In this case, even if not necessary, the investment is in any case desirable to guarantee the solvency of company.

1.2.1 General Aspects of Insurance Premium

The amount paid by the insured person, in relation to the random commitment taken by insurer with the stipulation of insurance contract, is defined as a premium.

The premium can be paid in a single payment at the time policy is signed, or it can be split into several installments. Most of non-life insurance contracts have an annual coverage, so premiums are generally paid annually by insured. The following pages will illustrate theoretical and practical aspects for calculation of various premium configurations of an insurance contract.

In the actuarial technique, there are different prize configurations: fair, pure, tariff or commercial and gross premium. The starting point is the fair premium, which corresponds to the expected value of the total random compensation paid by the insurance company during the insured period. Then there is pure premium, that includes the so-called security loading, which is the expected gain from the insurance contract for the company. This premium structure has the role of limiting any losses if the management of contract portfolio is negative due to estimation errors or an unexpected increase in claims. The next configuration is tariff premium, also known as a commercial premium. It is equal to the sum

between pure premium and fixed charges for expenses, intended to cover the costs of management and administration. This is therefore the premium that insurer asks for a coverage, but does not correspond to the price paid by insured. The latter is gross premium, which also has the taxes required by current regulations. Figure 1.2 shows the various premium configurations just presented.

$$\underbrace{\text{FAIR PREMIUM} + \text{Security Loading}}$$
$$\boxed{\text{PURE PREMIUM}}$$

$$\underbrace{\text{PURE PREMIUM} + \text{Fixed Loading}}$$
$$\boxed{\text{TARIFF PREMIUM}}$$

$$\underbrace{\text{TARIFF PREMIUM} + \text{Tax}}$$
$$\boxed{\text{GROSS PREMIUM}}$$

Figure 1.2 Representations of premiums

In the next sections will be explained in detail the first three types of premiums and how to calculate them. The gross premium will not be seen in detail because it is obtained by simply applying the taxes to the tariff premium.

1.2.2 Fair Premium

Calculation of the fair premium by theoretical approach

With the aim of identifying the fair premium, i.e. the premium that guarantees the correspondence between the amount paid by insured and the performance assumed by insurer, it is necessary to take into account the overall compensation payable by the insured in the year of coverage.

Let's considered a generic insurance contract, which provides for the coverage of the risk for a period of one year, let's indicate with:

- N the random number of claims during the coverage period. It is a natural and limited number. It may be difficult to determine the maximum number of claims, so often are considered as possible determinations of N all natural

numbers. In the choice of probability distribution, however, it must be considered that N is limited, choosing distributions that assign very low probabilities for high determinations of N;

- X_j the random amount of damage caused by the j-th claim;
- S total random compensation that is the sum of recompense for claims that affected the insured during the coverage period.

Since the insurer's objective is to estimate the random performance at the time the contract is signed, the random numbers N, X_j and S refer to the information held by the insurer at that precise moment. Therefore, fixed j, it is possible that the claim j-th is not realized and in this case X_j will be zero. If $N<j$ implies that $X_j=0$; if instead the j-th claim occurs, X_j is the amount of damage.

It is important to underline that only the claims that cause damage of a positive amount and that are considered compensable under the contractual conditions are relevant for the insurer. In fact it may happen that for some claims reported, the amount of damage charged to the insurer is null. These are the so-called *claims without follow-up*, which are not counted in the random number N.

To define S, compensation for individual damages is required: this is a function of the damage and will be indicated with Y_j, which indicates the random amount of compensation for the j-th claim and is defined by the function

$$Y_j = \phi(X_j)$$

where j=1,2,3... and ϕ indicates the compensation function, determined by the contractual conditions of the policy.

Assuming that the same policy conditions are applied to any claim, the random variables Y_1,...,Y_n are identically distributed and the total compensation for the damage S can be written as

$$S = \begin{cases} 0 & \text{if } N = 0 \\ Y_1 + ... + Y_N & \text{if } N > 0 \end{cases}$$

The *fair premium* P_f is given by

$$P_f = \mathbb{E}[S](1+i')^{-1/2}$$

where the expected value is calculated according to realistic assumptions for the claim amounts and the number of claim, while *i'* is the annual interest rate which expresses the time-value of money. In the following is assumed *i'=0*, since the duration of Car Liability is short.

The calculation of $\mathbb{E}[S]$ is usually performed accepting the following assumptions:

a. whatever the outcome n of N, the random variables Y_1, Y_2, ..., Y_n are stochastically independent and identically distributed;

b. the random number of claims is stochastically independent of the compensation Y_j, j=1,2,...,N.

Point b) implies that the probability distributions of the determinations of Y_j remains unchanged regardless of the hypothesis about N, then $\mathbb{E}[Y_j|N=n] = \mathbb{E}[Y_j]$ $\forall j$. However, a high determination of N should lead us to choose a distribution that assigns a very low probability for high determinations, since however, the number of claims that may occur during the insured year is limited. Therefore to very high values of N it will be appropriate to attribute low probability.

Indicating with Y a random variable with the same probability distribution as Y_j, from hypothesis a) it follows that:

$$\mathbb{E}[Y_j] = \mathbb{E}[Y]; j = 1,2,3,...$$

Imposing for n=1,2,...:

$$p_n = Pr(N=n)$$

the result is:

$$\mathbb{E}[S] = \sum_{n=0}^{\infty} p_n \mathbb{E}[S|N=n] = \sum_{n=1}^{\infty} p_n \mathbb{E}[S|N=n] =$$

$$\sum_{n=1}^{\infty} p_n \mathbb{E}[Y_1 + ... + Y_N|N=n] =$$

$$\sum_{n=1}^{\infty} p_n \sum_{n=1}^{\infty} p_n \mathbb{E}[Y_j | N = n].$$

Taking into account the two hypotheses set out above:

$$\sum_{n=1}^{\infty} p_n \sum_{n=1}^{\infty} p_n \mathbb{E}[Y_j | N = n] = \sum_{n=1}^{\infty} p_n n \mathbb{E}[Y] =$$

$$\mathbb{E}[Y] \sum_{n=1}^{\infty} p_n n = \mathbb{E}[Y]\mathbb{E}[N].$$

In conclusion, the expected value of global compensation is:

$$\mathbb{E}[S] = \mathbb{E}[Y]\mathbb{E}[N]. \tag{1.1}$$

Based on the assumptions made, for the definition of the expected value of the global reimbursement S, and therefore of the fair premium, all that the insurer needs is the estimate of the expected value of the number of claims $\mathbb{E}[N]$ and the expected value of the reimbursements $\mathbb{E}[Y]$ (Pitacco, 2000).

Calculation of the fair premium by statistical observation

To determine the fair premium it is possible to evaluate $\mathbb{E}[Y]$ and $\mathbb{E}[N]$ through the observation of a collective of insurance contracts, considering the number of claims that affected each contract, the relative damages and compensation.

The observation must be carried out on a group of contracts concerning risks similar to the one in question. This means that the policies must be similar in terms of type of assured risk, contractual conditions, possible amounts of compensation and so on.

Now suppose that a portfolio, consisting of a number r of contracts guaranteeing similar risks, is observed for a period of one year. It is assumed that the contracts were all stipulated at the same time and remained in the insurer's portfolio until expiry (each contract was therefore exposed one year to the possibility of being hit by one or more claims). It should also be assumed that, through the observation of this portfolio, the insurer has registered a total of m claims over the period of coverage, with compensation for amounts y_1, y_2, ..., y_m. It should be noted that the information refers to the portfolio as a whole, it is not

known which of the various contracts was affected by the claim. The ratio between the total compensation of the portfolio and the number of contracts, that is the *damage fee* per policy, is:

$$Q = \frac{y_1 + ... + y_m}{r}.$$

Q is the fair premium observed: in fact, if for each contract the payment of a premium equal to Q was requested, the insurer would have obtained the exact amount of compensation. Therefore entries and exits would be equal:

$$rQ = y_1 + ... + y_m.$$

Then the damage fee can be considered an estimate of the expected value of the compensation $\mathbb{E}[S]$. It is interesting to rewrite the Q as:

$$Q = \frac{m}{r}\frac{y_1 + ... + y_m}{m} = \frac{m}{r}\bar{y} \tag{1.2}$$

because the fair premium so decomposed allows to identify $\bar{y}$, which indicates the *average compensation* per claim, and $\frac{m}{r}$, that is the *claim experience index*, which reports the average number of claims by contract. This index may exceed the unit, as a contract may also be affected by more than one claim.

The (1.2) is the statistical image of the fair premium indicated in (1.1): the expected value of the single compensation $\mathbb{E}[Y]$ is estimated by $\bar{y}$, while the expected value of the number of claims $\mathbb{E}[N]$ is estimated through the use of the claims index $\frac{m}{r}$.

Regarding this index, if k is the maximum number of claims observed for a single contract, such that $k \leq m$, the number of contracts r of the portfolio can be divided as follows:

$$r = r_0 + r_1 + r_2 + ... + r_k$$

where r_h, with h=0,1,2,...,k, is the number of contracts affected by a number h of claims. The number of claims can be written as:

$$m = r_1 + 2r_2 + ... + kr_k.$$

It is then possible to decompose the claim index, as follows:

$$\frac{m}{r} = \frac{r_1 + 2r_2 + ... + kr_k}{r_1 + r_2 + ... + r_k}\frac{r - r_0}{r}$$

This decomposition makes it possible to identify two factors:

1. The *repeatability index*, defined by the first ratio, indicates the average number of claims relating to a damaged contract;

2. The second factor is the average number of claims for the single damaged contract, then indicates the average frequency of at least one claim. The quantity $\frac{r_0}{r}$, which is its complementary, indicates instead the frequency of non-claim.

It is important to note that, with the same degree of casualty index, a higher repeatability index indicates a higher concentration of claims on the same number of contracts. If this happens, the hypotheses of independence between claims and compensation, described in points a) and b) of the previous paragraph, must be reviewed through further study. In fact, a high concentration of claims in a small number of policies could be a sign of correlation between the various claims reported on a policy (Olivieri, Pitacco, 2010).

Exposure to the risk of heterogeneous portfolios

So far it has been assumed that the portfolio contracts are homogeneous with respect to the insured value, the stipulation date and the duration of coverage, thus simplifying the reality in some way.

However, the insurer usually has to pay compensation that comes from even very different exposures, based on the various insured values.

First of all, in order to compare the compensation for different insured values[2], it is necessary to eliminate the monetary dimension, i.e. to report all the quantities to an insured monetary unit. Indicating with w_1, w_2, ..., w_r the insured values or ceilings for each of the r observed contracts, it can be said that, the higher the value insured w_j in the policy, the higher the compensation the insurer expects to have to pay in the event of a claim. The *premium rate* can be measured as:

$$\tau = \frac{y_1 + y_2 + ... + y_m}{w_1 + w_2 + ... + w_r}.$$

[2]The assumptions for which the policies ensure the same type of risk, were stipulated at the same time and with the same annual duration, remain.

For policies that ensure higher values, the insured person will be required to pay a higher premium. In particular, the same premium rate, i.e. the same premium per unit of insured value, can be applied to all policies; they are in fact similar, apart from the insured value. The premium in this way will therefore be proportional to the insured value.

If τ is the premium rate applied, then the amount of receipts (premiums) for the insurer will coincide with the total exits (the compensation), with the same logic applied previously:

$$\tau(w_1 + w_2 + ... + w_r) = y_1 + y_2 + ... + y_m.$$

The average insured values is given by:

$$\bar{w} = \frac{w_1 + w_2 + ... + w_r}{r}.$$

representing the average exposure per contract. It is now possible to split τ as follows:

$$\tau = \frac{m}{r}\frac{\bar{y}}{\bar{w}} = \frac{Q}{w}.$$

The quantity indicated by $\frac{\bar{y}}{\bar{w}}$ is defined as *average claim degree*. As in the previous case where it was assumed that the portfolio was homogeneous, Q also expresses the average amount of compensation per policy in this case; however, due to the different insured values, such a piece of information is not appropriate neither for pricing, nor for summarizing the cost of claims incurred.

1.2.3 Pure Premium

The random financial transaction contemplated by the insurance contract can not be carried out solely in terms of equity; therefore a compensation for the risk that the insurer assumes must necessarily be added to the fair premium. This compensation is defined as safety loading and represents the expected gain for the company. In fact, if the insurer applied a fair premium to his policies, he would offer the contractor a disadvantageous contract, incurring the risk of suffering losses in the management of the contract portfolio due to the absence of any technical profit margin.

The pure premium, i.e. the sum between fair premium and security loading, is the premium that allows the insurer to achieve the technical balance of its

management, ensuring the solvency of the company and the ability to meet the commitments made towards the insured.

The pure premium can be defined as

$$\Pi = H(S) \tag{1.3}$$

The (1.3) is defined as *premium principle* and H indicates a functional that associates a real number Π with each possible probability distribution of the global compensation S.

Some mathematical properties should be satisfied by H, which are relevant from a practical point of view. The main properties are:

A. *Positivity of Safety Loading.*

It is necessary that the safety loading has positive value, for any S. The pure premium must be higher than the expected value of the compensation.

$$H(S) > \mathbb{E}[S]$$

B. *Additivity*

If S_1 and S_2 are two independent risks, is required that:

$$H(S_1 + S_2) = H(S_1) + H(S_2)$$

C. *Translation*

Given a positive real number b, it is necessary that:

$$H(S + b) = H(S) + b$$

The constant b represents an increase in the amount of compensation, common to all possible claims. If the possible amount of compensation S increases by a constant equal to b, then an equal amount increase will also be expected in the premium.

D. *Homogeneity of amount*

According to this property:

$$H(aS) = aH(S)$$

where a is a positive real number, which represents a proportional increase in the premium for each possible reimbursement. The homogeneity of the amount implies that the premium increases proportionally with the maximum amount that can be reimbursed.

However, this property is in contrast with the need to fix higher security loads for risks with a very high maximum reimbursement amount. In practice, the insurance company generally adopts an interim rate of premium at intervals, defined on the basis of the insured value. Within each interval the value of a will be constant, but increases as you go to successive intervals, then to higher assured values. The rate in this case will be constant at times; the homogeneity property will only be valid locally, within the individual intervals.

E. *Premium lower than the ceiling (No ripoff)*

If the compensation can not exceed a fixed amount M, called the ceiling, then:

$$H(S) \leq M$$

No insured person will be willing to pay a pure premium greater than the possible amount of compensation that he will realistically receive from the insurer in the event of a claim.

Moreover, there are five main principles for calculating the premium and are set out below.

I. *Expected Value Principle*

According to this principle, the pure premium is calculated as follows:

$$\Pi = (1 + \alpha)\mathbb{E}[S]$$

where $\alpha > 0$ is a given proportion and is dimensionless. The safeting loading, indicated by $\alpha\mathbb{E}[S]$, is proportional to the expected total payout of the insurer.

This is a calculation principle often used in insurance practice for its simplicity and because there is the advantage that the data required are the same used in the calculation of fair premium. However it has the disadvantage that the safety loading is not based on a risk measure.

II. *Variance Principle*

A safety loading proportional to a risk measure is instead foreseen by the principle of variance. In this case the pure premium is determined by the formula:

$$\Pi = \mathbb{E}[S] + \lambda Var(S)$$

where λ>0. It is noteworthy that $\lambda Var(S)$ must be an amount; since $Var(S)$ is an amount to the square, the dimension of λ must be that of $\frac{1}{amount}$. Otherwise said, λ is an intensity.

The safety loading $\lambda Var(S)$ is proportional to riskiness of the contract, measured by the variance. The ability of the safety loading to represent the expected gain for the enterprise depends on the ability of the variance to correctly quantify the risk deriving from S. To evaluate this, the probability distribution of S should be analysed: if it is "regular enough", i.e. it is symmetric and short tailed, then the variance is a good risk measure. Unlike the principle of expected value, the principle of variance therefore requires the analysis of additional information and data compared to those used for the calculation of the fair premium.

III. *Standard Deviation Principle*

Quite similar to the variance principle, the standard deviation principle (or average square deviation principle) assesses the pure premium as follows:

$$\Pi = \mathbb{E}[S] + \beta\sigma(S)$$

where β>0 is dimensionless and is a given proportion, while $\sigma(S) = \sqrt{Var(S)}$. The advantage compared to the variance principle consists in the fact that the parameter β is unit-free. Apart from this, the rationale of the two rules is similar; in particular, the same number for the pure premium could be determined under the two rules, provided that $\beta = \lambda\sigma(S)$.

IV. *Expected Utility Principle*

According to this principle, the pure premium is calculated as a solution to the equation:

$$\mathbb{E}[u(\Pi - S)] = 0$$

where u is the utility function. On the basis of this principle, the pure premium is that premium which makes the situation before the contract and the following one indifferent, in terms of expected utility: it is the minimum premium which makes the contract not disadvantageous for the insurer.

V. *Percentile Principle*

If $\Pi - S < 0$, i.e. $S > \Pi$, there is a situation of economic loss for the insurer. According to the percentile principle, the pure premium Π must be such that

$$Pr(S > \Pi) = \varepsilon$$

where $\varepsilon (0 < \varepsilon < 1)$ is the probability of achieving a loss on the individual contract and indicates a conveniently small percentile. The greater the probability of suffering a loss, the greater will be the pure premium.

The technical implementation of the previous rule may be time-consuming and clearly data for the estimate of the whole probability distribution of S are required. In practice, simpler rules are preferred, unless extreme risks are transferred to the insurer.

The principles of calculation of the pure premium do not satisfy all the listed properties of H, but only some of them. Figure 1.2 summarizes which of the properties are satisfied by the individual principles.

	I	II	III	IV	V
A	✓	✓	✓	✓	X
B	✓	✓	X	X	X
C	X	✓	✓	✓	✓
D	✓	X	✓	X	✓
E	X	X	X	✓	✓

Figure 1.2: The property satisfied by the principles of premium calculation.

Property A is satisfied by all the principles, with the exception of the percentile. If the probability of occurrence of the claim is lower than ε, then the premium will be 0. It is therefore necessary to verify that the safety loading is

positive, in order to avoid that the defined premium is lower than the expected value of compensation ($\Pi < \mathbb{E}[S]$), hypothesis that would lead to the failure of the company.

As for the property of additivity, it is respected by the principle of expected value, because the expected value of a sum is equal to the sum of the individual expected values, and the principle of variance, since the variance of a sum of independent random variables corresponds to the sum of their variances. The property is not valid for the principle of standard deviation, expected utility and percentile. In the first case it is not valid because the standard deviation of the sum of independent random variables, generally is not equal to the sum of their respective standard deviations; it does not even apply to the percentile principle given that the percentile of a distribution is not equal to the sum of percentiles of the individual probability distributions. In the case of the expected utility principle, the property is not valid if the utility is quadratic, but if it is considered an exponential utility[3], this property is satisfied.

Property C finds application in all principles, except in principle I because $H(S+b) = (1+\alpha)(\mathbb{E}[S]+b)$ is not equal to $H(S)+b = (1+\alpha)\mathbb{E}[S]+b$.

It is then observed that the property of the homogeneity of amount is not valid for the principle of variance and expected utility, both in the case of quadratic utility and in the case of exponential utility.

Finally, property E is only satisfied by the expected utility and percentile principle. This property is not applicable to the variance principle because if two risks are taken, S and $T{=}zS$ (with $z{>}0$), will be obtained $M_T = zM_S$, where M indicates the maximum possible compensation. If the principle is applied to identify the premium related to T risk, it will be obtained $\Pi_T = z\mathbb{E}[S] + Bz^2Var(S)$. From the equation, the premium Π_T will be higher than the maximum possible compensation M_S if $z > \frac{M_S - \mathbb{E}[S]}{BVar(S)}$. Not even for the principle of the average square deviation the property is satisfied, because if it puts

$$S = \begin{cases} 0 & \text{with probability } q \\ 1 & \text{with probability } p = 1-q \end{cases}$$

[3]The exponential utility formula is given by: $u(s) = B(1 - e^{\frac{S}{B}})$

then $M_S = 1$, the application of the principle will lead to a premium of $\Pi = p + \beta\sqrt{pq} > 1$ if $p > \frac{1}{1+\beta^2}$.

1.2.4 Tariff Premium

The tariff premium, or commercial premium, which will be indicated with P_T, is the premium requested from the contractor, which takes into account the expenses charged to the company. The tariff premium is calculated by adding so-called *loading for expenses* to the pure premium. The classes of expenses are essentially of three types:

- costs for contract acquisition;
- costs for collection of premium;
- costs for administrative management.

Acquisition costs refer to the phase of stipulation of the contract or, in any case, to expenses attributable to the first year of coverage. They include the purchase commissions, the costs of issuing the policy and the costs for any medical examinations or investigations.

The collection costs relate to collection commissions, receipt rights and accounting for collections; they are incurred in correspondence with the payment of the installment of the premium.

Management costs are expenses not directly attributable to the individual contract, in fact each policy is assigned a quota calculated on the basis of a fixed percentage of the insured capital.

Acquisition and collection costs are commensurate in terms of a α percentage of the P_T, while for management costs, the costs are recovered by applying a β percentage, again on the tariff premium.

Indicating with Π the pure premium, the tariff premium will be:

$$P_T = \Pi + (\alpha + \beta)P_T$$

and then

$$P_T = \frac{1}{1 - (\alpha + \beta)}\Pi$$

The loading coefficients α and β are different according to the risk branch and depend on the volume of the portfolio and on the market conditions.

So far, it has been assumed that the tariff premium was paid by the contractor in a single solution at the time the contract was signed. However, it is frequent that it is expected to be divided into several installments, the number of which is indicated with k: if k=2 means that the installment is semi-annual, if k=3 the installment is quarterly and so on. The installment P_T^k is necessarily higher than the fraction $\frac{1}{k}P_T$ of the annual premium. This is because it is essential to take into account the increase in management charges and the loss of interest.

A coverage of less than one year may also be requested; in this case a proportional premium reduction will not be applied because the management and acquisition costs are not reduced proportionally to the duration of the policy.

1.3 Tariff Process

The tariff process leads the insurer to determine the premium to be applied to policyholders and, as seen in the previous paragraphs of this chapter, the premium is defined on the basis of the probabilistic assessment of the insurer's provision or total compensation due in consequence of the claims, which affected the insured risks in the period covered by the policy. Under the hypothesis of a composite distribution of total compensation, the objective is to determine the technical basis for each risk, that is to assign the distributions of the number of claims and damage per claim through the use of data deriving from the observation of the portfolio of the insurer, assuming that the claim occurs.

The insurance portfolios are made up of a heterogeneous set of risks and this heterogeneity is due to endogenous factors, inherent in the particular nature of the risk, but also to exogenous factors typically environmental or socio-economic. Through the pricing techniques, the insurer divides the risk community into subgroups or classes, which have similar characteristics, so as to be able to attribute the same technical basis to the risks belonging to the same class. Through this process, the premiums are therefore differentiated for the insured, depending on the different risk profile.

This differentiation of premiums takes place in two phases. In the first phase,

called *a priori personalization or pricing*, the premiums are differentiated according to the characteristics of the risks, observable at the moment of the conclusion of the contract, before having any kind of information on the claims of the insured deriving from the experience. The insurer identifies sub-groups of similar risks, called *tariff classes or risk classes*, based on *tariff variables*, and then assesses the premiums to be attributed to each class. Despite the use of a large number of tariff variables, within the classes there remains, however, heterogeneity for the behavior of the insured and the claims. In order to try to formulate forecasts of the insured person's claims, the observation of his "insurance history" may be more effective, so much so that for some coverage it is expected an adjustment of the premium in the aftermath. This is therefore the second phase of the charging process: *a posteriori customization of premium*. This takes into account the experience on the claims of each insured acquired during the period covered by the policy. In this way a change from a collective premium for each class to a premium based on individual experience will be implemented (Gigante, Picech, Sigalotti, 2010).

1.3.1 Risk Classes and Personalization a Priori.

To determine the characteristics of the risks on the basis of which to personalize the a priori premiums, reference is made to statistical observations of company data, portfolios of other companies or even market data. Thanks to the statistical observation, the insurer can determine which factors influence the probabilistic variations of the random elements that describe the claim experience of each individual.

The claim experience can be described by considering the number of claims incurred during the coverage period, the compensation for the claim or the total compensation. Other elements may also be considered, for example, in the case of civil liability it may be important to separate the damage caused by accidents to property, injury to persons or damage to both.

The definition of the risk classes is therefore based on:

- *the risk factors*, i.e. the characteristics of the risk. Factors judged influential on the claims experience. It is information that the insurer can obtain a

priori, before having data on the history of claims, such as the sex of the insured or the power of car;

- *the modalities*, qualitative or quantitative, that represent the determinations assumed by the risk factors, such as "female" (qualitative) or "69 CV" (quantitative).

The choice of risk factors and the definition of the modalities are based on the use of statistical methods such as cluster analysis, with the aim of dividing the modalities of the explanatory variables into classes, or univariate or multivariate statistical analysis procedures, to define the ordering of the explanatory variables on the basis of their significance. In the case of univariate analysis, it refers to the chi-square test, while in the case of multivariate analysis are used the generalized linear models and the random forest.

Once the insurer has identified the risk factors, it is necessary to group the modalities to select the most significant factors. The risk must then be assigned to an appropriate class, so as to be able to assign each insured person to the class that most reflects his claim experience; the classes in which the portfolio is divided are called *tariff classes* and the risk factors selected are called *tariff variables*.

In order to evaluate the premiums, it is necessary to identify a function, the so-called *tariff model*, which allows the corresponding premium to be associated to each class, thanks to some parameters on which it depends, called *relativity*. Once the tariff model has been chosen and relativity is estimated, it is possible to obtain the tariff.

In the next chapter, generalized linear model will be examined, that will be used to estimate the value of premium.

Chapter 2

Generalized Linear Models

For the pricing processes, linear regression models may not be fully suitable, in particular for what concerns the pricing of the non-life insurance. First, the number of claims follows a discrete probability distribution and the damage amounts caused by the claims have as support the positive half-straight line, generally with a positive asymmetric distribution. Moreover, in many cases the hypothesis of a linear link between the expected value of the response variable and the determinations of the explanatory variables is not acceptable. For this reason, as mentioned in the previous chapter, an extension of the linear regression models is used for the a priori definition of the tariff: Generalized Linear Model (GLM).

The GLMs allow to overcome some limits that characterize the linear models since, instead of assuming a normal distribution, they assign to the variables response distributions belonging to the linear exponential class, which includes in addition to the normal, for example, also the Poisson distribution (typically attributed to the number of claims), and the Gamma distribution (frequently adopted for the distribution of the amount of damage).

2.1 Theory of Generalized Linear Models

Considering a number n of statistical units, the set of corresponding observations will be $\{y_i, \mathbf{x}_i, i = 1,2,...,n\}$[1]. For each statistical unit i, y_i indicates the value

[1]It is specified that the letters or symbols indicated in bold are matrices. While with the symbol " ’ " during the chapter will be indicated the transposition.

of a quantity of interest and the various y_i form the vector $\mathbf{y}$, which indicates the observed value of the aleatory vector of the response variables $\mathbf{Y}$. Instead, $\mathbf{x}_i$ is the vector of the determinations assumed by the explanatory variables (also called *covariates*) considered. Through the use of GLM, the distribution of $\mathbf{Y}$ and the vector of the determinations assumed by the covariates are related.

There are two types of hypotheses, probabilistic and structural, that define GLMs:

- it is assumed that the response variables $Y_1, ..., Y_n$ are stochastically independent and that the respective distributions all belong to a specific linear exponential family;
- it is hypothesized the presence of a link between the expected value of Y_i, indicated with μ_i, and the vector x_i. The link is expressed by:

  $$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

  where $\boldsymbol{\beta}$ is a vector of parameters, g is the *link function* and is invertible and $\mathbf{x}_i'\boldsymbol{\beta}$ is the *linear predictor*. Then the expected value of Y_i will be calculated as:

  $$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

Therefore for the definition of a GLM several elements must be taken into consideration: the distribution of the response variable vector, the explanatory variables, the linear predictor, the link function and the parameters.

2.1.1 The Distribution of Response Variables Vector

As previously said, for the response variables, indicated with $Y_1, ..., Y_n$, it is assumed that there is stochastic independence and that they belong to the same *linear exponential family.*

A linear exponential family is a parametric family of non-degenerate[2] probability distribution, whose density function can be written as:

$$f(y; \theta, \phi) = exp\{\frac{y\theta - b(\theta)}{\phi}\}c(y, \phi)$$

[2]A non-degenerate distribution is a probability distribution not concentrated in a single value coinciding with the average.

where θ and ϕ are two real parameters, while b and c are real functions. In particular, c is the normalization, while b is defined *cumulant function* and characterizes, within the class of linear exponential distributions, a determined distribution family; the parameter θ is called *canonical parameter*, while ϕ is the *dispersion parameter*.

The distributions that are part of the class of linear exponential families and the elements that characterize them are summarized in Figure 2.1.

	θ	ϕ	b(θ)	c(y,ϕ)
Gaussian N(μ,σ^2)	μ	σ^2	$\frac{\theta^2}{2}$	$(2\pi\phi)^{-1/2}exp\{-\frac{y^2}{2\phi}\}$
Poisson P(μ)	logμ	1	e^θ	$\frac{1}{y!}$
Scaled Binomial B(n,π)/n	$\log(\frac{\pi}{1-\pi})$	$\frac{1}{n}$	$log(1+e^\theta)$	$\binom{1/\phi}{y/\phi}$
Negative Binomial BN(μ,α)	$\log(\frac{\mu}{\alpha-\mu})$	1	-αlog$(1-e^\theta)$	$\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}$
Gamma G(α, μ)	-$\frac{1}{\mu}$	$\frac{1}{\alpha}$	-log(-θ)	$\frac{1}{\phi}^{\frac{1}{\phi}}(\frac{y^{\frac{1}{\phi}-1}}{\Gamma(\frac{1}{\phi})})$
Inverse Gaussian GI(μ,β)	$-(2\mu^2)^{-1}$	β	$-(-2\theta)^{\frac{1}{2}}$	$(2\pi\phi y^3)^{-\frac{1}{2}}exp\{-\frac{1}{2\phi y}\}$

Figure 2.1: Families of the linear exponential class.

For what concerns the expected value and the variance of linear exponential families, it can be said that:

$$\mathbb{E}[Y] = \mu = b'(\theta)$$

and

$$Var(Y) = \phi b''(\theta)$$

where b' and b'' indicate respectively the first derivative and the second derivative of the cumulant function. The variance formula shows that random variables do not have the property of homoskedasticity, i.e. they do not all have the same variance, as in the case of a linear regression model, but they are heteroskedastic.

After having exposed the main characteristics of the linear exponential family, the probability or density function of Y_i can be defined as:

$$f(y;\theta_i,\phi,\omega_i) = exp\{\frac{\omega_i}{\phi}[y\theta_i - b(\theta_i)]\}c(y,\phi,\omega_i)$$

in which θ_i and ϕ are the canonical and dispersion parameters respectively, ω_i is an assigned weight greater than 0 and b and c are the two real functions. It is to be noted that the cumulant function b does not change with respect to i because

the linear exponential family has been fixed; the one that changes according to i is the canonical parameter.

In relation to the moments of the linear exponential class distribution:

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$$

and

$$Var(Y_i) = \frac{\phi}{\omega_i} b''(\theta_i)$$

There is an important observation to make: taking into account the weights, with the same dispersion parameter, the variance of Y_i will be greater the lower the ω_i weight.

2.1.2 The Explanatory Variables

The explanatory variables represent observable characteristics, which influence the probabilistic evaluation of the response variables. There are two types of explanatory variables:

1. *numerical variables*, which have numerical determinations, such as the variable "age of the insured";

2. *categorical variables*, which have non-numerical determinations, such as the variable "sex of the insured".

Variables that have a numerical determination can be entered directly into the model, while non-numeric variables must be previously numerically encoded, through a binary variable.

According to the general rule, a classification variable with l modes can be coded with indicator variables, called *dummy variables*. Considering a categorical variable C , whose modalities are indicated with c_1, c_2, ... , c_l, the variable can be expressed with the l indicator variables:

$$X = \begin{cases} 1 & \text{if C} = c_i \\ 0 & \text{otherwise} \end{cases}$$

with $i = 1, 2, \ldots, l$.

Considering that $\sum_{i=1}^{l} X_i = 1$, *l - 1* indicator variables will be sufficient to describe C and the remainder can be deduced by complement to 1. After transforming all the categorical variables into numerical ones, the observed characteristics can be represented by m numerical variables X_1, X_2, ... , X_m, whose determinations for the i-th observation will be indicated with x_{i1}, x_{i2}, ... , x_{im}.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}$$

The matrix $\mathbf{X}$ of type n x p, where $p = m + 1$, is defined *regression matrix*, in which the first column consists of unitary elements and the *(j + 1)* column shows the determinations of variable X_j for each observation i. $\mathbf{X}$ therefore includes all the determinations of the explanatory variables observed.

In the continuation of the chapter let us assume that $n > p$ and that $\mathbf{X}$ is with full rank p; this means that the columns will be linearly independent.

2.1.3 The Linear Predictor

If x_{i1}, ... , x_{im} are the determinations of explanatory variables and β_0, ... , β_m are parameters common to all statistical units, the linear combination $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} = \mathbf{x}_i'\boldsymbol{\beta}$ is called linear predictor. It can also be written in matrix form as $\eta = \mathbf{X}\boldsymbol{\beta}$ and represents the systematic component of the model, according to the parameters β_j, which are not known, but estimated.

2.1.4 Link Function

The link function g is a real invertible function, which relates the elements of the linear predictor η_i to the expected value of the response variable μ_i in this way

$$\eta_i = g(\mu_i)$$

Considering the formula of linear predictor expressed in the previous paragraph, it is obtained

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

The link function is considered a monotonic function in the strict sense that admits the first and second continue derivatives.

In the pricing process, g indicates how to calculate the fair premium according to the characteristics of the tariff variables, therefore through g the tariff model is determined:

- if g is an *identical* link function, i.e. $g(\mu) = \mu$, it is obtained $\mu_i = \mathbf{x}_i'\boldsymbol{\beta} = \sum_{j=0}^{m} x_{ij}\beta_j$ and then an additive tariff model;
- if g is an *logarithmic* link function, i.e. $g(\mu) = log(\mu)$, it is obtained $\mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}} = \prod_{j=0}^{m} e^{x_{ij}\beta_j}$, arriving at a multiplicative tariff model;
- if g is an *power* link function, i.e.

$$g(\mu) = \begin{cases} \frac{\mu^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ log(\mu) & \text{if } \gamma = 0 \end{cases}$$

  for $\gamma = 1$ it is obtained $g(\mu) = \mu - 1$ and therefore return to the identical link function; while regarding $\gamma \to 0$, $g(\mu) \to log(\mu)$. So if γ varies between 0 and 1, there is a change from an additive model to a multiplicative model;

- if g is a *canonical* link function,

$$g(\mu_i) = b'^{-1}(\mu)$$

  and in this case

$$\eta_i = g(\mu_i) = \theta_i$$

  with $i = 1, 2, ..., n$. This function varies according to the distribution considered within the linear exponential family, as indicated by Figure 2.2.

2.1.5 Parameters

Finally, as regards the parameters, it has already been mentioned that two types of parameters are involved in GLM: the canonical parameters θ_1, θ_2, ..., θ_n and the dispersion parameter ϕ. These parameters are not normally known, but can be estimated using data.

Distribution family	b(θ)	b'(θ)	$g(\mu) = b'^{-1}(\mu)$
Gaussian	$\frac{\theta^2}{2}$	θ	μ
Poisson	e^{θ}	e^{θ}	$log(\mu)$
Scaled Binomial	$log(1+e^{\theta})$	$\frac{e^{\theta}}{e^{\theta}+1}$	$log(\frac{\mu}{1-\mu})$
Negative Binomial	-αlog$(1-e^{\theta})$	$\alpha\frac{e^{\theta}}{1-e^{\theta}}$	$log(\frac{\mu}{\alpha+\mu})$
Gamma	-log(-θ)	-$\frac{1}{\theta}$	-$\frac{1}{\mu}$
Inverse Gaussian	$-(-2\theta)^{\frac{1}{2}}$	$-(-2\theta)^{-\frac{1}{2}}$	$-\frac{1}{2\mu^2}$

Figure 2.2: Canonical link functions.

For what concerns the estimation of canonical parameters, it is carried out by estimating the parameter vector $\boldsymbol{\beta}$ (which will be estimated in turn): in fact, once estimated $\boldsymbol{\beta}$, will be get the determination of θ_1, θ_2, ..., θ_n through the formula

$$\theta_i = b'^{-1}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))$$

In the particular case where the canonical link function was chosen, the canonical parameters will be given by

$$\theta_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$$

with $i = 1, 2, ..., n$.

As mentioned above, to estimate the canonical parameters it is sufficient to estimate the vector of the regression parameters $\boldsymbol{\beta}$ and for the estimation of this vector the most used method is that of maximum likelihood, in particular the log-likelihood.

Remembering that $\mathbf{y} = (y_1, y_2, ..., y_n)$' is the observed value of the vector of the response variables $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$', the log-likelihood l of $\mathbf{y}$ is:

$$l(\theta, \phi; y) = logL(\theta, \phi; y) = \sum_{i=1}^{n} \{\frac{\omega_i}{\phi}[y_i\theta_i - b(\theta_i)] + logc(y_i, \phi, \omega_i)\}$$

$$= \sum_{i=1}^{n} l_i(\theta_i, \phi; y_i)$$

Each canonical parameter can be expressed as a function of $\boldsymbol{\beta}$, so the log-likelihood can be indicated as $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta})$. The maximum likelihood estimates are obtained by identifying the relative maximum points of l from the system of likelihood equations given by

$$\left\{ \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = 0 \quad \text{j} = 0, ..., \text{m} \right.$$

The resolution of this system is not the focus of this thesis, so will not be seen how to get to the solution $\hat{\boldsymbol{\beta}}$.

Also the dispersion parameter ϕ, when not known, must be estimated and also in this case the method used is that of maximum likelihood, but this time the equation to be solved is

$$\left\{ \frac{\partial l(\hat{\theta},\phi;y)}{\partial \phi} = 0 \right.$$

where $\hat{\theta}$ indicates the estimate of the vector of canonical parameters, obtained from $\hat{\boldsymbol{\beta}}$.

For the purpose of determining the fair premium to be applied to the various tariff classes, the insurance companies generally estimate the distribution of the number of claims and damage per claim separately.

2.2 Models for the Number of Claims

Regarding the distribution of the number of claims, the GLMs aim to estimate the probability distribution of the number of claims affecting the insured or the tariff class within a year.

For the construction of the models, in this paragraph it is assumed that the explanatory variables, the number of claims occurring during the observation period and the exposure, i.e. the duration measured in years of the contractual coverage period, are available.

The number of tariff classes into which the portfolio is divided will be indicated with K and with reference to the tariff class k, with $k = 1, ..., K$, will be indicated with x_k the vector of determinations of the explanatory variables common to the risks of the class and with n_k the number of insured persons within that tariff class. Moreover, for the i-th insured in class k, m_{ki} will indicate the number of claims occurring during the observation period to the insured and will represent the observed value of the random number of claims M_{ki}, with $i = 1, ..., n_k$, which affect the risk insured during the observed period and which are supposed to be stochastically independent. Finally, t_{ki} will indicate the exposure, $\eta_k = \mathbf{x}_k'\boldsymbol{\beta}$ the linear predictor and $\lambda_k > 0$ the expected annual number of claims for each

policyholder of class k.

Therefore:

$$g(\lambda_k) = \eta_k$$

$$\lambda_k = g^{-1}(\eta_k)$$

with $k = 1, ..., K$.

The model most commonly used to describe the number of claims is the *Poisson*[3] distribution, indicated as:

$$M_{ki} \sim P(t_{ki}\lambda_k)$$

The parameter of the distribution is the product between the exposure and the expected number of claims for the insured of class k.

Based on the Poisson distribution indicated above, the probability density function will be:

$$Pr(M_{ki} = y) = e^{-t_{ki}\lambda_k}\frac{(t_{ki}\lambda_k)^y}{y!} = e^{-t_{ki}e^{\theta_k}}e^{y\theta_k}\frac{(t_{ki})^y}{y!} =$$

$$= e^{y\theta_k - t_{ki}e^{\theta_k}}\frac{(t_{ki})^y}{y!}$$

with $y = 0, 1, ...$ and $\theta_k = log(\lambda_k)$.

On the basis of this distribution, it is possible to construct a GLM for the estimation of the frequencies of claims for the classes or for single insured.

2.2.1 GLM for Individual Claims Frequencies

For this model, it is necessary to consider for each policy of the portfolio the frequency of the claims of the insured, calculated as:

$$Y_{ki} = \frac{M_{ki}}{t_{ki}}$$

with $k = 1, ..., K$ and $i = 1, ..., n_k$. Based on the probability density function:

$$Pr(Y_{ki} = y) = Pr(M_{ki} = t_{ki}y) =$$

$$= e^{t_{ki}y\theta_k - t_{ki}e^{\theta_k}}\frac{(t_{ki})^{t_{ki}y}}{(t_{ki}y)!} = e^{t_{ki}(y\theta_k - e^{\theta_k})}c(y, t_{ki})$$

[3]Alternatively, the negative binomial distribution $M_{ki} \sim BN(\alpha, t_{ki}\lambda_k)$ can be used.

with $y = 0, t_{ki}^{-1}, 2t_{ki}^{-1}, ...$

The probability distribution of Y_{ki} then follows the distribution of the linear exponential family with cumulative function $b(\theta) = e^{\theta}$, canonical parameter θ_k, dispersion parameter $\phi = 1$ and weight $\omega_{ki} = t_{ki}$. The expected value and the variance of Y_{ki} will instead be given by:

$$E[Y_{ki}] = e^{\theta_k} = \lambda_k$$

$$Var(Y_{ki}) = \frac{1}{t_{ki}}\lambda_k$$

The expected value, since $\lambda_k = g^{-1}(\eta_k)$, can be rewritten as $E[Y_{ki}] = g^{-1}(\eta_k)$. At this point, a GLM can be built for the frequencies of individual claims Y_{ki}, with $k = 1, ..., K$ and $i = 1, ..., n_k$. In the GLM:

- the response variables Y_{ki} are stochastically independent and follow a Poisson distribution with weight t_{ki}, $E[Y_{ki}] = \lambda_k$ and $Var(Y_{ki}) = \frac{1}{t_{ki}}\lambda_k$;
- the linear predictors are indicated by $\eta_k = \mathbf{x}_k'\boldsymbol{\beta}$;
- the link function is g. If the canonical function for the Poisson distribution, which is the logarithm, was chosen as g, this would lead to the definition of a multiplicative model, which is the one most used in insurance practice.

2.2.2 GLM for Claims Frequencies in Classes

In this model the frequency of the claims in the tariff class Y_k is considered, which can be calculated as a weighted average of the individual frequencies, in which the weights are the exposures. Therefore will have:

$$Y_k = \sum_{i=1}^{n_k} \frac{t_{ki}}{t_k} Y_{ki} = \frac{M_k}{t_k}$$

where $t_k = \sum_{i=1}^{n_k} t_{ki}$ is the total exposure of the class k and $M_k = \sum_{i=1}^{n_k} M_{ki}$ is the total number of claims in the same class k.

In the same way as in the previous case, the distribution of Y_k is found and is

$$Pr(Y_k = y) = e^{t_k(y\theta_k - e^{\theta_k})} c(y, t_{ki})$$

with $y = 0, t_{ki}^{-1}, 2t_{ki}^{-1}, ...$

The frequency of claims therefore has a Poisson distribution with weight t_k and

$$E[Y_k] = e^{\theta_k} = \lambda_k$$
$$Var(Y_k) = \frac{1}{t_k}\lambda_k$$

Choosing, like the model for the individual claims frequencies, the logarithm function as a link function, a model with data grouped with a structure equivalent to the previous model is obtained.

The GLM built has the following structure:

- the response variables Y_k are stochastically independent and follow a Poisson distribution with weight t_k, $E[Y_k] = \lambda_k$ and $Var(Y_k) = \frac{1}{t_k}\lambda_k$;
- the linear predictors are indicated by $\eta_k = \mathbf{x}_k'\boldsymbol{\beta}$;
- the link function is g.

As can be seen, this case is very similar to the previous one.

2.2.3 GLM for Numbers of Individual Claims

It is possible to define a GLM also for the number of claims. Setting $\mu_{ki} = t_{ki}\lambda_k$ and $log(\mu_{ki}) = \theta_{ki}$, the probability density function for the number of individual claims will be:

$$Pr(M_{ki} = y) = e^{-\mu_{ki}}\frac{(\mu_{ki})^y}{y!} = e^{y\theta_{ki}-e^{\theta_{ki}}}\frac{(t_{ki})^y}{y!}$$

with $y = 0, 1, ...$, $k = 1, ..., K$ and $i = 1, ..., n_k$.

As required in the definition of GLM, M_{ki} follows a distribution belonging to linear exponential family, but the expected value poses some problems: in fact, since $g(\lambda_k) = \eta_k$, the expected value will be:

$$\mu_{ki} = E[M_{ki}] = t_{ki}\lambda_k = t_{ki}g^{-1}(\eta_k)$$

The link between the expected value and the linear predictor is not the one required in the definition of the GLMs, unless the exposures of the policies are all unitary. Choosing the logarithm as link function g, then the expected value can be rewritten as:

$$\mu_{ki} = t_{ki}e^{\eta_k} = e^{lnt_{ki}+\eta_k} = g^{-1}(lnt_{ki} + \eta_k)$$

In this way the term lnt_{ki}, called *offset*, was introduced, thanks to which a GLM for the number of claims can be built. The term offset therefore introduces an explanatory variable with a known effect in the linear predictor and can be used to place a constraint on some regression parameters, requiring that they assume fixed values; in this case the constraint is imposed on the exposures.

The GLM will have the following structure:

- the response variables M_{ki} are stochastically independent with Poisson distribution, $E[M_{ki}] = Var(M_{ki}) = \mu_{ki} = t_{ki}\lambda_k$;
- the linear predictors are indicated by $\eta_k = lnt_{ki} + \mathbf{x}'_k\boldsymbol{\beta}$, where lnt_{ki} is the offset term;
- the link function g is the logarithm.

2.2.4 GLM for the Number of Claims in the Classes

In this last model, the total number of claims is considered for each tariff class instead of for a single insured person.

The total number of claims is indicated with M_k, while t_k represents the total exposure, with $k = 1, ..., K$. Considering that $M_k = \sum_{i=1}^{n_k} M_{ki}$, starting from the initial hypothesis according to which $M_{ki} \sim P(t_{ki}\lambda_k)$ and the stochastic independence hypothesis of M_{ki}, it derives that also M_k is a Poisson, in particular $M_k \sim P(t_k\lambda_k)$. Setting $\mu_k = t_k\lambda_k$ and $ln(\mu_k) = \theta_k$, the probability function for M_k will be:

$$Pr(M_k = y) = e^{-\mu_k}\frac{(\mu_k)^y}{y!} = e^{y\theta_k - e^{\theta_k}}\frac{1}{y!}$$

with $y = 0, 1, ...$ and $k = 1, ..., K$.

Similarly to before, since $g(\lambda_k) = \eta_k$, the link between expected value and linear predictor is not the one required in the GLMs because

$$\mu_k = E[M_k] = t_k\lambda_k = t_k g^{-1}(\eta_k)$$

Also in this case, if g is the logarithm, the insertion of offset term lnt_k occurs and a GLM with the following structure is obtained:

- the response variables M_k are stochastically independent with Poisson distribution, $E[M_k] = Var(M_k) = \mu_k = t_k\lambda_k$;

- the linear predictors are indicated by $\eta_k = lnt_k + \mathbf{x}_k'\boldsymbol{\beta}$, where lnt_k is the offset term;

- the link function g is the logarithm.

2.3 Models for Damage per Claim

The objective of the damage claim model is to estimate the probability distribution of the damage per claim in a specific tariff class. To do this, it is essential to assume that the explanatory variables have been set and to have the data regarding the determinations of the explanatory variables, the number of claims and the amounts of damage.

It is important to note that some problems are typically found on the damage amounts. In fact, first of all, the presence of limits to compensation, such as maximum amount, insurance deductibles or overdraft, mean that the data recorded by the insurance companies are those of the compensation amount and not the effective amount of damage. Another problem is related to the claims whose amount of damage has only been estimated, so the amount is not yet a certain value; the estimate is generally obtained by examining any partial payments already made or the values entered in the reserve. The problem is that the number of claimed policies is very small, so there are few data available; it is even more difficult to estimate the high damage amounts because most of these are small amounts.

These problems often make the estimates deriving from the models for damages less reliable than the estimates relating to the number of claims.

As regards the definition of the models, the number of tariff classes into which the portfolio is divided will be indicated with K, while the specific tariff class with $k = 1, ..., K$. $\mathbf{x}_k'$ will indicate the vector of determinations of the explanatory variables and $\eta_k = \mathbf{x}_k'\boldsymbol{\beta}$ the linear predictor. Finally, m_k indicates the total number of claims that hit a given tariff class and ns_k the number of policies in the class, which suffered at least one claim.

The most used distribution for the determination of the damage per claim is

the *Gamma* distribution. The cumulative function of the Gamma family, considering that $b(\theta) = -ln(-\theta)$, is:

$$f(y; \theta_k, \phi) = e^{\frac{1}{\phi}[y\theta_k + ln(-\theta)]} c(y, \phi)$$

Moreover, the expected value μ_k is linked to the linear predictor through the link function g and it will be:

$$g(\mu_k) = \eta_k$$

$$\mu_k = g^{-1}(\eta_k)$$

with $k = 1, ..., K$.

Based on this distribution, it is possible to build several GLMs.

2.3.1 GLM for Damages per Claim

The random damage caused by the i-th claim for the risks of the tariff class k is Y_{ki} and it is assumed that these random numbers are stochastically independent, with $k = 1, ..., K$ and $i = 1, ..., m_k$.

The probability distribution, under these hypotheses, will be:

$$Y_{ki} \sim e^{\frac{1}{\phi}[y\theta_k + ln(-\theta)]} c(y, \phi)$$

with

$$E[Y_{ki}] = \mu_k = g^{-1}(\eta_k)$$

$$Var(Y_{ki}) = \phi {\mu_k}^2$$

It is therefore obtained a GLM with:

- the response variables Y_{ki} represent damage per claim, are stochastically independent and follow a Gamma distribution with $E[Y_{ki}] = \mu_k$ and $Var(Y_{ki}) = \phi {\mu_k}^2$;
- the linear predictors are indicated by $\eta_k = \mathbf{x}_k' \boldsymbol{\beta}$;
- the link function is $g = -ln(-\theta)$.

2.3.2 GLM for Average Damage per Claim in the Classes

In a model like this, the response variable of each tariff class corresponds to the arithmetic average of the damage per claim in the class, i.e.

$$Y_k = \frac{1}{m_k} \sum_{i=1}^{m_k} Y_{ki}$$

with $k = 1, ..., K$.

The Gamma probability density function will then be:

$$Y_k \sim e^{\frac{m_k}{\phi}[y\theta_k + ln(-\theta)]} c(y, \phi, m_k)$$

Compared to the previous case, here there is the addition of the weight m_k. The expected value and variance of Y_k will be

$$E[Y_k] = \mu_k = g^{-1}(\eta_k)$$

$$Var(Y_k) = \frac{\phi}{m_k} {\mu_k}^2$$

The structure of this GLM is defined by:

- the response variables Y_k are stochastically independent and follow a Gamma distribution with weight m_k, expected value $E[Y_{ki}] = \mu_k$ and variance $Var(Y_{ki}) = \frac{\phi}{m_k} {\mu_k}^2$;
- the linear predictors are indicated by $\eta_k = \mathbf{x}_k' \boldsymbol{\beta}$;
- the link function is $g = -ln(-\theta)$.

2.3.3 GLM for Average Damage per Claim for Accident Policies

Supposing to have only the data relating to the total damages for the accident policies, the models described above cannot be used.

For the j-th accident insurance belonging to the tariff class k, C_{kj} will indicate the total random damage and m_{kj} the number of claims in this class k, with $k = 1, ..., K$ and $j = 1, ..., ns_k$. The general observation relates to an insured who has suffered at least one claim, therefore $m_{kj} \geq 1$.

The average damage per claim for the j-th accident insurance of class k can be calculated as:

$$\bar{Y}_{kj} = \frac{C_{kj}}{m_{kj}}$$

The distribution of $\bar{Y}_{kj}$ will be

$$\bar{Y}_{kj} \sim e^{\frac{m_{kj}}{\phi}[y\theta_k + ln(-\theta)]} c(y, \phi, m_{kj})$$

This is always a Gamma distribution, but in this case the weight used is the total number of claims for accident policies m_{kj}. Then

$$E[\bar{Y}_{kj}] = \mu_k = g^{-1}(\eta_k)$$

$$Var(\bar{Y}_{kj}) = \frac{\phi}{m_{kj}} {\mu_k}^2$$

This GLM is built on the basis of:

- response variables $\bar{Y}_{kj}$ stochastically independent, which follow a Gamma distribution with weight $m_k j$, expected value $E[\bar{Y}_{kj}] = \mu_k$ and variance $Var(\bar{Y}_{kj}) = \frac{\phi}{m_{kj}} {\mu_k}^2$;
- linear predictors indicated with $\eta_k = \mathbf{x}_k' \boldsymbol{\beta}$;
- link function $g = -ln(-\theta)$.

2.3.4 GLM for Average Damage per Claim of Accident Policies in the Classes

In this case the response variable indicates the average damage per claim in the tariff class k, considering only damaged policies, and is:

$$Y_k = \sum_{j=1}^{ns_k} \frac{m_{kj}}{m_k} \bar{Y}_{kj} = \sum_{j=1}^{ns_k} \frac{m_{kj}}{m_k} \frac{C_{kj}}{m_{kj}} = \frac{1}{m_k} \sum_{j=1}^{ns_k} C_{kj}$$

The distribution of Y_k is

$$Y_k \sim e^{\frac{m_k}{\phi}[y\theta_k + ln(-\theta)]} c(y, \phi, m_k)$$

with

$$E[Y_k] = \mu_k = g^{-1}(\eta_k)$$

$$Var(Y_k) = \frac{\phi}{m_k} {\mu_k}^2$$

The structure of GLM for average claims damages in the various tariff classes consists of:

- response variables Y_k stochastically independent, with Gamma distribution and weight m_k, expected value $E[Y_k] = \mu_k$ and variance $Var(Y_k) = \frac{\phi}{m_k}{\mu_k}^2$;
- linear predictors indicated with $\eta_k = \mathbf{x}'_k\boldsymbol{\beta}$;
- link function $g = -ln(-\theta)$.

The model can be considered adequate when the empirical coefficients of variation[4] are more or less constant to vary of tariff classes. If the damage amounts per claim are known, the empirical coefficients of variation for the individual tariff classes are:

$$cv_k = \frac{\sqrt{\frac{1}{m_k - 1}\sum_{i=1}^{m_k}\left(y_{ki} - \bar{y}_k\right)^2}}{\bar{y}_k}$$

where $\bar{y}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} y_{ki}$.

As regards the non-empirical variation coefficients, for the class k the expected value is μ_k and the variance $\phi{\mu_k}^2$, therefore the coefficient of variation will be:

$$\frac{\sqrt{\phi{\mu_k}^2}}{\mu_k} = \sqrt{\phi}$$

One possibility of improving the model, in case of unsatisfactory adaptation, is the introduction of a weight ω_k, which represents the weight for the distribution of damage per claim in class k. The coefficient of variation will therefore be rewritten as:

$$\frac{\sqrt{\frac{\phi}{\omega_k}\mu_k^2}}{\mu_k} = \sqrt{\frac{\phi}{\omega_k}}$$

In the next chapter the models described will be applied to a practical case.

[4]The coefficient of variation is by definition the ratio between standard deviation and average.

Chapter 3

Empirical Application

In the previous chapters, an attempt was made to provide a general overview of the non-life insurance products and the methods for defining the premium, through the use of generalized linear models, while in this chapter a practical case will be discussed.

In the first part of this chapter some useful Machine Learning techniques will be explained for the construction of model, in the second part the work dataset will be presented and finally the third part will be dedicated to the process for the definition of premium. In particular, generalized linear models will be applied to determine a premium *a priori*, with the approach generally used by insurance companies, which provides for the separate estimate of the frequency of claims and amounts of damage.

The sample of policies on which the analysis will be based was kindly provided by the RGI company, an Independent Software Vendor specialized in the Insurance Industry. The company has 800 professionals and 12 offices spread across the EMEA[1] region (Italy, Ireland, France, Germany, Luxembourg and Tunisia). Its main product is PASS, a modular Policy Administration System that enables the end-to-end management of Policies, Claims and Insurance Products configuration across all the insurance channels and business lines. RGI is a leader of its sector in the European market with 103 installations for the insurance companies and other 300 for the brokers.

[1]EMEA is the acronym of Europe, Middle East and Africa and is a geographical designation used mainly in the economic-industrial field.

3.1 Useful Machine Learning Techniques

3.1.1 Decision Tree

The first technique considered is the *Decision Tree*, which will not be used directly in the model, but it is the basis of Random Forest which will be used.

Decision Tree is a supervised[2] learning algorithm, used to visually and explicity represent decisions. As the names goes, it uses a tree-like model of decisions: it is composed by nodes, which represent a certain value, and in each node data are splitted in two o more child nodes, with the purpose to separate labels in the best way possible. The initial node, from which everything starts, is called the *root*, while a *leaf* is a child node that no longer needs to be split, because all elements, or at least the great majority, have the same label. The variable that best separates the elements is selected from a certain number of candidates, settable with the '*mtry*' command in R.

To check if the data has been split well, it is necessary to define an index, called *Gini Index*, that can be calculated for each node with the formula:

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

where c is the number of different classes and p_i is the proportion of elements of class i in the examined node. Defining n_i the number of elements of class i, p_i can be obtained as:

$$p_i = \frac{n_i}{\sum_{j=1}^{c} n_j}$$

The variable that best splits the generic node k is the one that maximizes the $GiniIndexDecreaseD_k$, calculated as the difference between the Gini Index of the node k and the weighted sum of the Gini Index of its child nodes, i.e.:

$$D_k = Gini_k - (\frac{N_1}{N_1 + N_2} Gini_{k_1} + \frac{N_2}{N_1 + N_2} Gini_{k_2})$$

where N_1 and N_2 are respectively the number of elements in the first and second child node.

[2]The goal of supervised algorithm is to find specific relationships or structure in the input data that allow to effectively produce correct output data, using explicitly-provided labels (training set).

Another parameter of trees is the *maximum depth*, settable in R with the command '*maxdepth*', that imposes a maximum number of consequent splittings for the tree. Practically, it refers to the length of the longest path from a root to a leaf. It is advisable to set this parameter to have no huge trees, which involve high complexity and possible overfitting.

A very simple example of decision tree is present below.

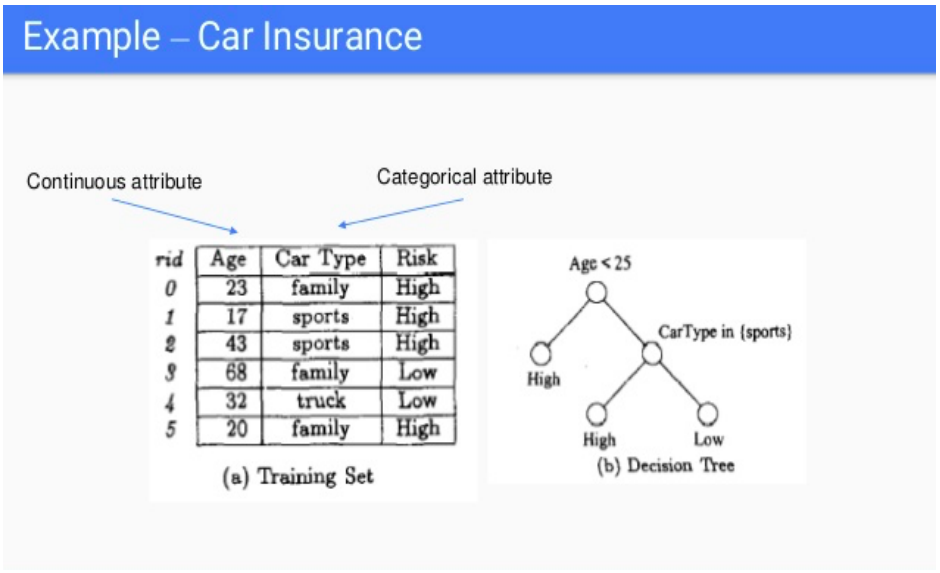

Figure 3.1: Example of decision tree in the insurance field.

3.1.2 Random Forest

The second technique dealt with is the *Random Forest*, that will be useful to see which are the most significant risk factors for the model: in fact, a quality of the Random Forest algorithm is that it is very easy to measure the importance of each feature on the prediction. Observing the variables importance, it is possible to decide which features to keep and which to delete, since they don't contribute enough or nothing to the prediction process, in order to reduce the computational cost.

Random Forest is, like the Decision Tree, a supervised learning algorithm and it is used both for classification and regression[3]. As the name suggests, this algorithm builds and merges multiple decision trees together to get a more

[3]Problems with a quantitative response are *regression* problems, while those involving a qualitative response are *classification* problems.

accurate and stable prediction and it adds additional randomness to the model, while growing the trees. In fact, instead of looking for the most important variable while a node is split, it searches for the best feature among a random subset of variables. Creating random subsets of features and building smaller trees using these subsets, Random Forest prevents overfitting most of the time.

The subtrees are then combined and this can make the computation slower, based on how many trees the Random Forest builds. This is why hyperparameters are used: to increase the predictive power of the model and to make the model faster.

In R the main hyperparameters are:

- *ntree*: is the number of trees. A more accurate prediction requires more trees, but a large number of trees can make the algorithm slow and ineffective for real-time predictions. It is therefore necessary to choose a number that is not too large, so that the prediction is good;

- *maxnodes*: represents the maximum number of leaves in the forest. If not given, trees grow as much as possible and a warning is issued if it is set larger than the maximum possible;

- *mtry*: similarly to the hyperparameter of the Decision Tree, it represents the number of variables randomly sampled as candidates at each split. It should be noted that in R the default values are different for classification ($\sqrt{p}$, where p is the number of variables in dataset) and regression ($\frac{p}{3}$).

Random Forest is considered an algorithm simple and handy since its default parameters often produce a good prediction. In this thesis however, it will not be used for this property, but because it is easy to view the relative importance to the input features.

3.1.3 Balancing Techniques

Downsampling and *Oversampling* are approaches used when the dataset is unbalanced: in particular, they are used when the variable response to predict has very disproportionate classes inside. For example, in this thesis, a response variable to predict is the frequency of a claim and the available data are 97,26% related

to non-claims and only 2,8% to claims. It is therefore necessary to balance the two classes.

Downsampling produces a subset of the initial dataset, which contains all the elements of the minority class and a randomly selected subset of the majority class, so that the two classes are balanced. Instead, oversampling duplicates elements of the minority class, up to having a number equal to the majority class.

Data balancing techniques are very powerful tools, but they must be used carefully, in order to get a correct evaluation of model's performances. If they are not applied correctly, the model's ability of class recognition can be estimated incorrectly. First of all, downsampling and oversampling must be applied only on the training set, while model obtained must be tested on a dataset where classes have the original unbalanced ratio. The same approach must be maintained in the application of cross-validation[4]: data must be divided in training and validation set before balacing (in R this division is made automatically with the command *'train'* of the package *'caret'*).

As an example, in Figure 3.2 is shown a wrong application of oversampling, i.e. when duplicates of the minority class are generated before the execution of cross-validation. The problem is that the same elements appears in both validation and training set, causing overfitting.

Instead, in Figure 3.3 is shown the correct application of the algorithm, i.e. when data duplication is made after the creation of validation set, in order to avoid any kind of overfitting.

In this thesis, to balance the dataset, downsampling technique will be used. The disadvantage is that, as the majority class is reduced, some data is lost. However, there is the advantage of reducing the computational cost: in fact, if oversampling was used, having lots of data available, creating duplicates would lead to very large dataset sizes and therefore a very high computational cost.

[4]Cross-validation is a validation technique for assessing how the results of a model will generalize to an independent dataset.

Figure 3.2: Wrong application of oversampling during cross-validation.

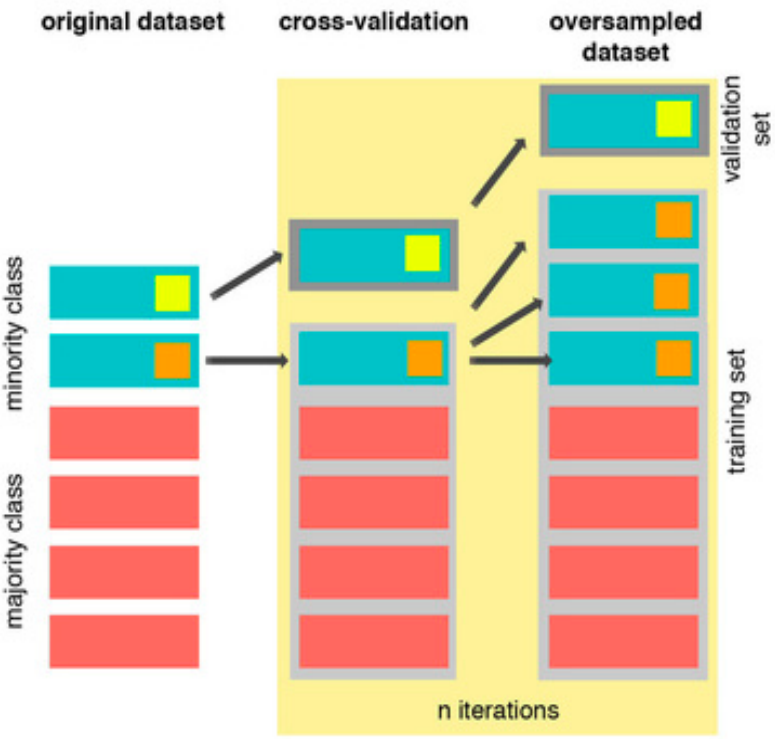

Figure 3.3: Correct application of oversampling during cross-validation.

3.2 Description of Model Data

3.2.1 Presentation of Datasets

The data available concern the non-life insurance, in particular policies relating to Italian RCA in 2018, and are divided into two datasets:

1. Dataset relating to *claims*: it consists of 25466 observations and for each of these there are 11 variables:

   - *Id*: it is an integer that indicates what position a policy has within the given dataset;
   - *Chiave univoca polizza*: it represents uniquely the policy (even outside the reference dataset);
   - *Numero pratica*: it is the number of the insurance practice in which the policy is present;
   - *Rischio colpito*: it indicates what kind of accident happened. For the analysis present in this thesis only claims related to the RCA will be selected;
   - *Data denuncia* and *Data evento*: respectively, date and time when the claim was reported and when it occurred;
   - *Flag chiuso*: it is 1 if the practice has been closed, 0 otherwise;
   - *Data chiusura*: date and time when the insurance practice ended. It is present only in correspondence with *Flag chiuso = 1*;
   - *Sinistri costo totale*: it represents the total cost of the claim to be paid;
   - *Sinistri liquidazioni pagato*: sum of money that has been liquidated by insurance. When the practice is closed, this value is certainly different from zero;
   - *Sinistri riserve passive*: it is a prudent global assessment of debts to third parties for claims that have already occurred (due to the insurer's liability), but not yet paid. Unlike before, if the practice is closed this value is equal to zero.

2. Dataset relating to *portfolio*: it contains 1048575 rows and 27 columns:

   - *Id* and *Chiave univoca polizza*: equal to the previous dataset;
   - *ID Polizza*: it is another number that uniquely represents the policy;
   - *Operazione*: it can be *"Sostituenti"*, if it is an insurance contract that replaces a previous policy, *"Nuovi"* otherwise;
   - *Frazionamento*: it is the partition of commission and it can be *"Annuale"* or *"Semestrale"*;
   - *Data emissione*, *Data effetto* and *Data scadenza*: respectively represent the date on which the policy is issued, on which it begins to be valid and on which it has no more effect;
   - *ID Contraente*: integer that uniquely distinguishes the various policyholders;
   - *Data nascita*, *CAP Contraente* and *Sesso Contraente*: date of birth, ZIP code and sex of insured;
   - *Marca*, *Modello*, *Settore veicolo*, *Classe veicolo* and *Uso veicolo*: various information relating to the insured car. The most important information is the type of car (sport, luxury, utility car, etc.);
   - *Valore Assicurato*: value of the insured object;
   - *CU di origine* and *CU Assicurata*: respectively previous and current class of merit;
   - *Rischio assicurato*: it is the type of risk insured. Only the RCA will be considered in this thesis;
   - *Massimale RCA Cose*, *Massimale RCA Persone* and *Massimale RCA Sinistro*: respectively they are the maximum for damage to property, to people and to the claim in general. As mentioned in Chapter 1, the contractor may decide to have a single maximum amount per claim (therefore *Massimale RCA Persone* and *Massimale RCA Cose* are equal to *Massimale RCA Sinistro*) or two separate maximum amounts (*Massimale RCA Sinistro* is equal to the sum of the other two);

- *Indice Credit Scoring (rischio)*: credit risk index relating to the policyholder. Possible values are from 0 to 5: 0 indicates that the risk is low, while 5 indicates that the risk is high.;
- *Premio Netto*: net premium of the insurance in question;
- *Agenzia*: place where the insurance agency is located.

3.2.2 Data Preparation

Both datasets require a data cleaning and preparation process before applying the Generalized Linear Models and this can be divided into the following steps.

Variables and Rows Reduction

First of all, it is advisable to eliminate all the variables and the rows that are not necessary for the models, in order to reduce computational time and costs.

Regarding the first dataset (*DatasetSinistri*), the unnecessary variables are *"Id"* and *"Numero.pratica"*, because *"Chiave.univoca.polizza"* is enough to uniquely identify the policies.

For the second dataset (*DatasetCosti*), it is necessary to eliminate more variables: *"Id"*, *"ID.Polizza"*, *"Data.emissione"*, *"ID.Contraente"*, *"Settore.veicolo"*, *"Classe.veicolo"*, *"Uso.veicolo"* and *"CU.origine"*.

Instead with regards to the rows, for both datasets only those corresponding to the RCA risk must be selected.

New Variable Creation

The variables of the datasets do not always represent at best a characteristic that is needed for the analysis and therefore it is necessary to create new ones through those that are available.

For example, it is necessary to know the exposure of the policy that is not present in the data. However, it can be calculated dividing by 365 the period between the start of the effect and the expiry of the policy.

Another necessary variable is the age of the insured, obtainable by making the difference between a vector containing the date 31/12/2018 and the vector containing the dates of birth of the various contractors.

The variables used to create the features are no longer necessary, so they can be deleted from the datasets.

Elimination of Outliers

A very important step for data cleaning is the elimination of outliers, i.e. the anomalous values present in the datasets. These data must be removed because they are very different from other data and they can be errors that compromise the analysis, leading to an unrealistic result.

Since these are datasets relating to the RCA, it must be checked that all the insured people are over 18 years old. In fact, all minors cannot take out car insurance and therefore they are outliers.

Furthermore, all ZIP codes must be exactly 5 digits, so all ZIP codes with different length from this one must not be taken into account for analysis.

It is also necessary to check the value assumed by *"CU.Assicurata"*, because the universal merit classes must be different from 0 and from numbers greater than 18.

Finally, net premiums with a negative value or less than 16€ are eliminated.

Elimination of NaN

The last step in data cleaning consists in eliminating Not A Number (NaN), which do not allow the correct operation of some machine learning algorithms that will be applied later. In general, the NaN indicate missing values and they can be replaced by the average of the available values, for example. However in this case, they are simply not taken into account since, even if those data are eliminated, a lot of information is available.

Once these data preparation steps have been carried out, the two datasets can be merged using the primary key *"Chiave.univoca.polizza"*, present in both, through the *left-join* command of R. This function returns all rows from the left table *DatasetCosti* and any rows with matching keys from the right table *DatasetSinistri*. The resultant data frame *DatasetNew* is composed by 128039 observations and it is the dataset on which the analysis will be made.

3.3 Exploratory Analysis

Before applying the model, it is necessary to make an exploratory analysis of the data to see their behaviour and to obtain useful information. This analysis is made on the variables that are considered to be the most significant for the future model. For each of these, it is necessary to understand what is the frequency and the average cost of claims.

Age of Insured

To find the frequency of claims based on the age of insured, only the contractors who have made a claim must be considered and they must be divided into eight groups based on age. The number of claims corresponding to each age group is then calculated.

```
NClaimsAge1 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age <= 21))
NClaimsAge2 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 21 & Age <= 25))
NClaimsAge3 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 25 & Age <= 35))
NClaimsAge4 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 35 & Age <= 45))
NClaimsAge5 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 45 & Age <= 55))
NClaimsAge6 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 55 & Age <= 65))
NClaimsAge7 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 65 & Age <= 75))
NClaimsAge8 <- nrow(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 75))
```

For example, the *NClaimsAge1* variable contains the total number of claims made by insured persons aged between 18 and 21 years.

However, these numbers must be normalized with respect to the total exposure corresponding to the respective age group.

```
ExpAge1 <- sum(filter(DatasetNew, Age <= 21)$EsposizioneTot)
ExpAge2 <- sum(filter(DatasetNew, Age > 21 & Age <= 25)$EsposizioneTot)
ExpAge3 <- sum(filter(DatasetNew, Age > 25 & Age <= 35)$EsposizioneTot)
ExpAge4 <- sum(filter(DatasetNew, Age > 35 & Age <= 45)$EsposizioneTot)
ExpAge5 <- sum(filter(DatasetNew, Age > 45 & Age <= 55)$EsposizioneTot)
ExpAge6 <- sum(filter(DatasetNew, Age > 55 & Age <= 65)$EsposizioneTot)
ExpAge7 <- sum(filter(DatasetNew, Age > 65 & Age <= 75)$EsposizioneTot)
ExpAge8 <- sum(filter(DatasetNew, Age > 75)$EsposizioneTot)
```

ExpAge1 is the sum of exposures corresponding to all the policies (with and without claims) relating to insured persons who are between 18 and 21 years old. To normalize, *NClaimsAge1* must be divided by *ExpAge1* and the same must be done for the other age groups.

Regarding the average cost, instead, the first thing to do is to find the total costs of the claims relating to each age group.

```
CostTotAge1 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age <= 21)$Sinistri.Costo.Totale)
CostTotAge2 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 21 & Age <= 25)$Sinistri.Costo.Totale)
CostTotAge3 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 25 & Age <= 35)$Sinistri.Costo.Totale)
CostTotAge4 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 35 & Age <= 45)$Sinistri.Costo.Totale)
CostTotAge5 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 45 & Age <= 55)$Sinistri.Costo.Totale)
CostTotAge6 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 55 & Age <= 65)$Sinistri.Costo.Totale)
CostTotAge7 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 65 & Age <= 75)$Sinistri.Costo.Totale)
CostTotAge8 <- sum(filter(na.omit(DatasetNew, cols = Sinistri.Costo.Totale), Age > 75)$Sinistri.Costo.Totale)
```

Once the total cost has been found, it is sufficient to divide this by the number of claims (previously found).

The final graphs, obtained with the R *barplot* command, are:

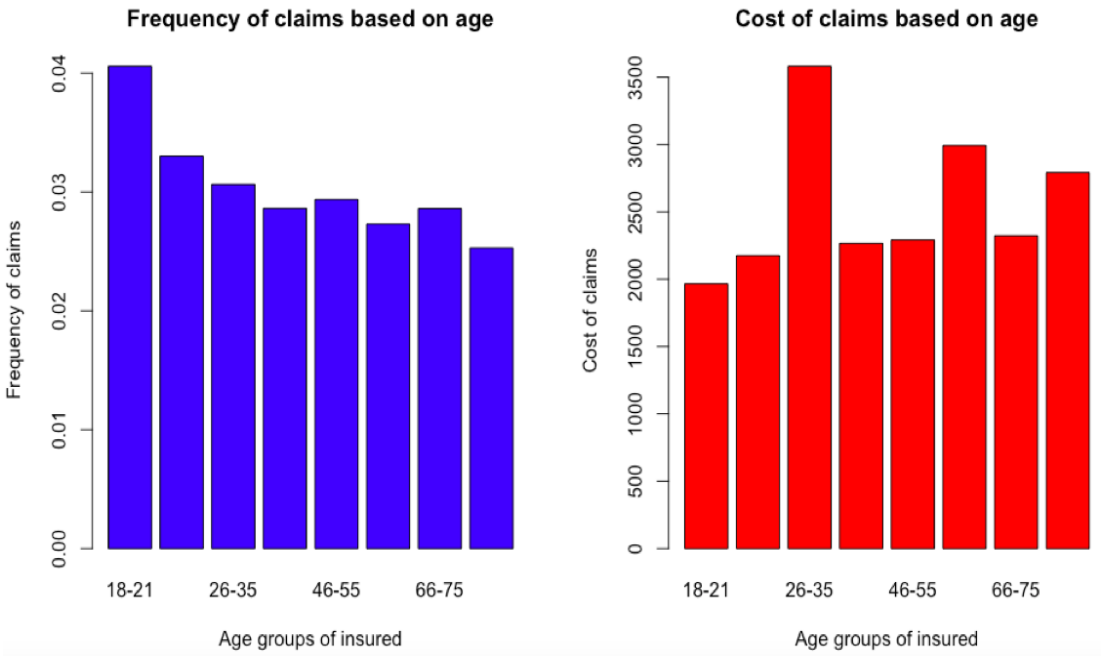

Figure 3.4: Histograms of frequency and average cost of claims based on age.

From Figure 3.4 it is clear that, as regards the frequency of claims, at the 18-21 range, in which the majority of new drivers are included, there is a peak and then there is a decreasing trend. With regard to the average cost, however, the trend is increasing (except for a peak in the 26-35 range), because probably at the beginning that one has less experience, cars with lower value are used.

Sex of Insured

Even for the sex of insured, the basic idea is the same as before. The insured are divided into two classes, male and female, and the frequency and average cost of the claims are calculated for each of the two categories, using the previous method.

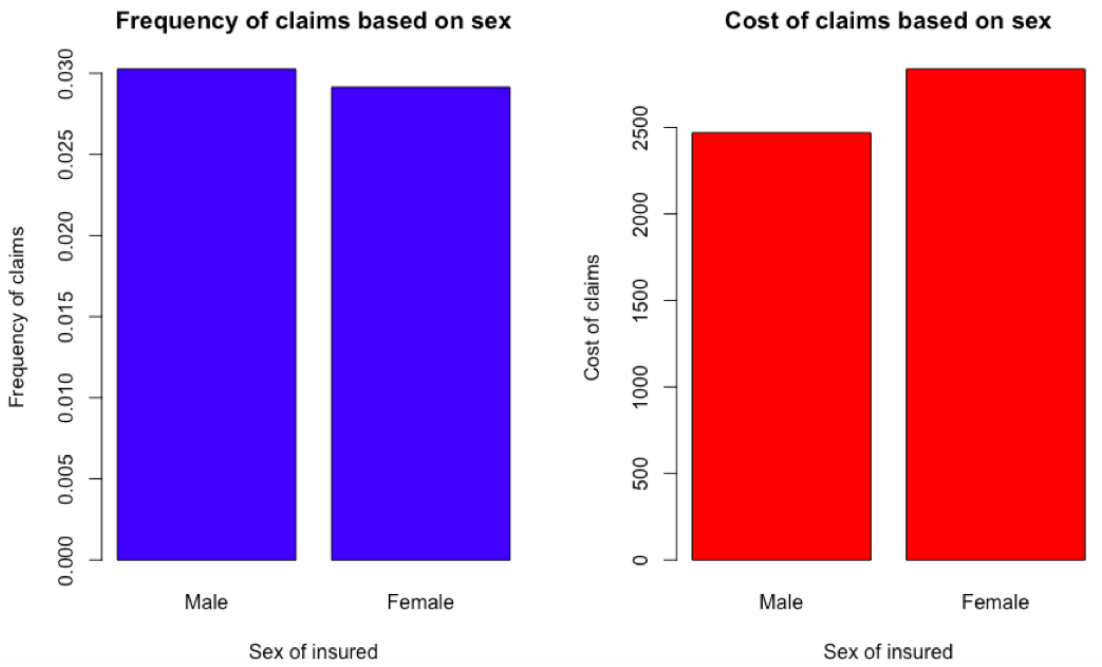

Figure 3.5: Histograms of frequency and average cost of claims based on sex.

Observing Figure 3.5, it should be noted that the frequency of claims and the average cost do not differ much from male or female.

ZIP Code of Insured

For this exploratory analysis, Italy is divided into four zones (North, Central, South and Islands), based on the first two digits of the ZIP code. For the subdivision, Figure 3.6 was taken as a reference.

The northern area includes ZIP codes ranging from 10 to 46, the central zone those from 00 to 06 and from 47 to 67, the south from 68 to 89 and finally the islands include codes from 07 to 09 and from 90 to 98.

The graphs obtained based on the above subdivision are shown in Figure 3.7.

The most claimed areas are those of central Italy and the islands, but there is not much difference compared to the other two areas. However, in the islands there is a lower average cost of claims than in the rest of Italy.

Universal Merit Class

It was decided to group the merit classes. In total there are six groups: the first group (containing classes from 1 to 3) contains the insurances relating to insured who make fewer claims, while the sixth group (which has the classes from 16 to

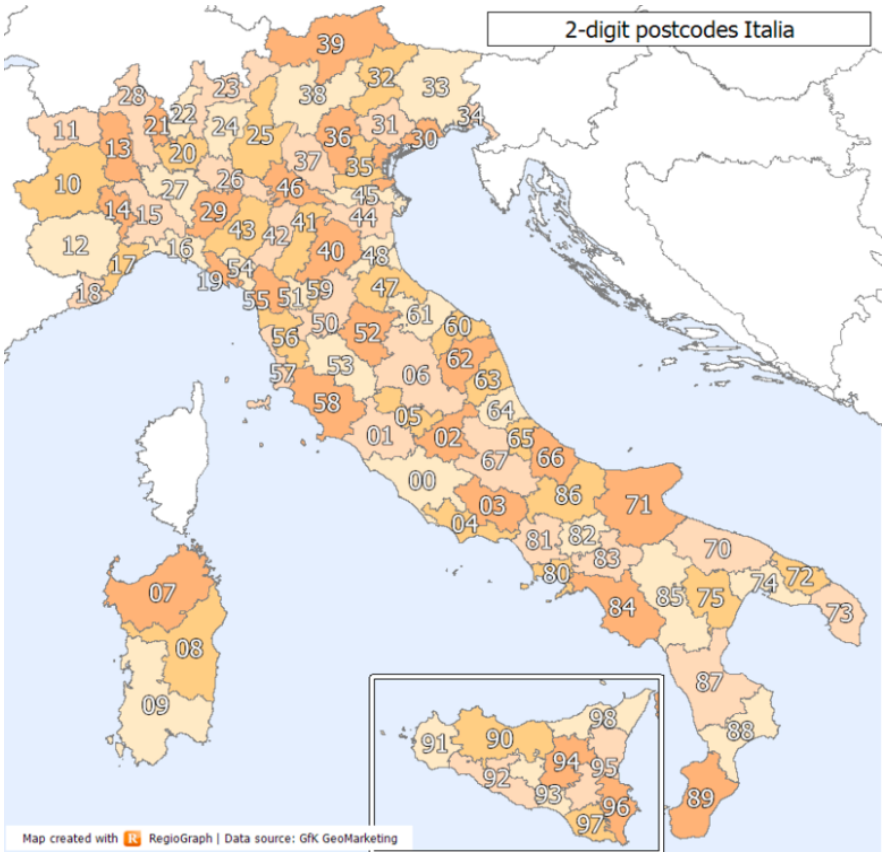

Figure 3.6: Subdivision of Italy based on the first two digits of the ZIP codes.

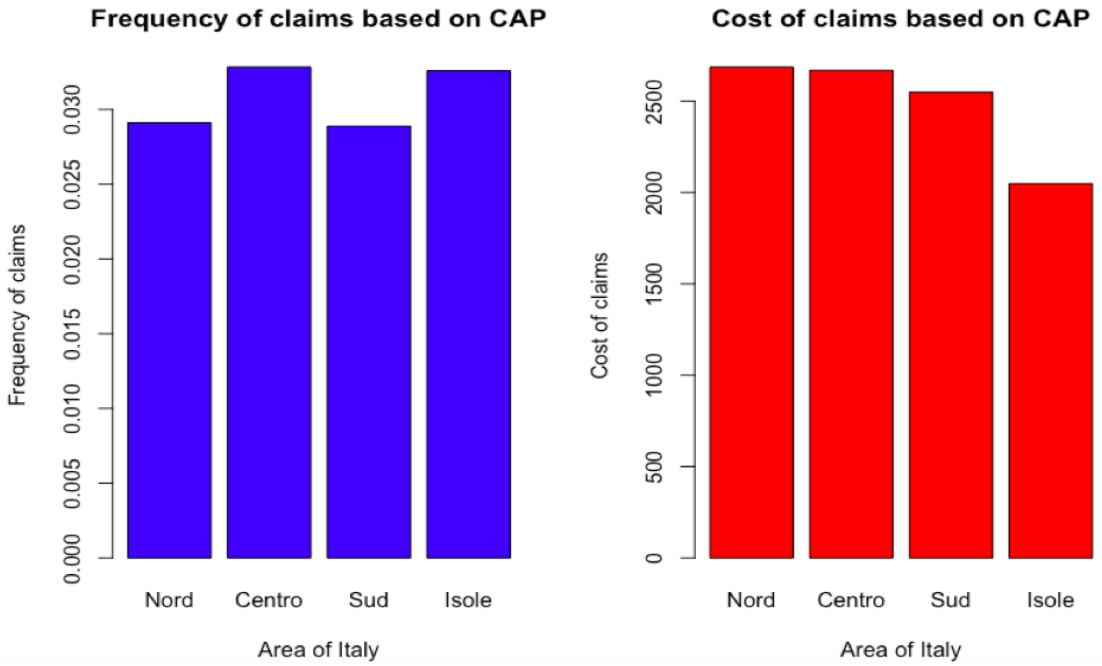

Figure 3.7: Histograms of frequency and average cost of claims based on ZIP code.

18) concerns policies of the contractors who have made more claims.

As for the previous variables, to obtain histograms, the number of claims

normalized for the total exposure and the total cost divided by the number of claims are used. The result is Figure 3.8.

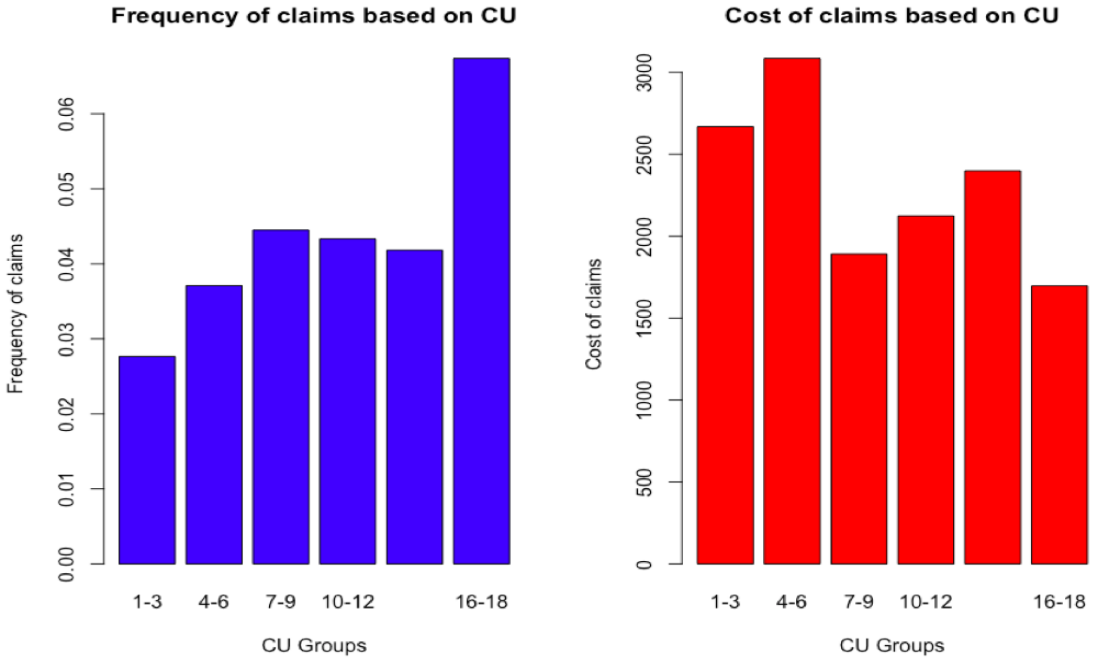

Figure 3.8: Histograms of frequency and average cost of claims based on merit classes.

For the reason explained above, trend in the frequency of claims is increasing with a peak in the sixth group. Average cost, instead, seems to be greater for the first two groups, which contain within them the six highest classes of merit.

Car brand

The last variable taken into consideration in this exploratory analysis is the car brand. According to the brand, three groups were created: luxury, medium and normal category.

The first includes all insurance related to luxury cars, so the average cost for this is expected to be higher than the other two. On the contrary, normal category should be the one with the lowest average cost.

The decreasing trend of the second histogram of Figure 3.9 confirms the one described above, while from the first histogram it can be seen that most of the claims are made with cars of the middle category.

From the exploratory analysis made, it can be concluded that the variables

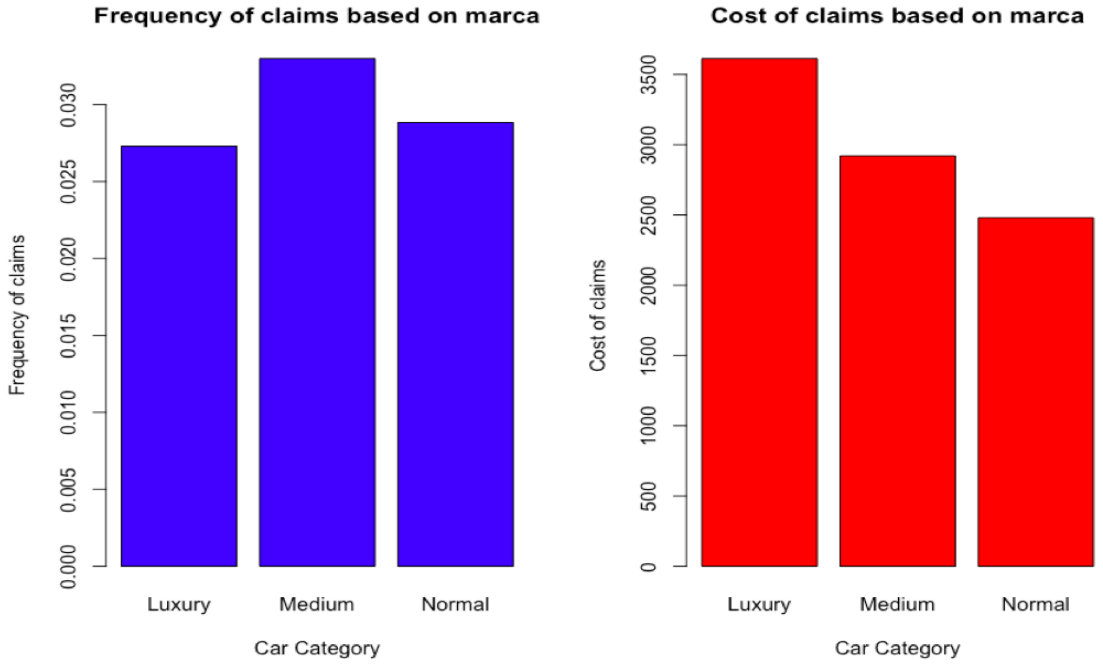

Figure 3.9: Histograms of frequency and average cost of claims based on car brand.

that seem to be most significant for the number of claims are age and class of merit. For the average cost, instead, the car category also appears to be significant.

However, this exploratory analysis based on graphs is not sufficient to exclude the less significant variables. It is better to use a more "objective" method to understand the importance of variables: the Random Forest.

3.4 Selection of the Most Important Variables

To reduce the computational cost, it is advisable that GLMs have as variables only those that influence more the probabilistic evaluation of the response variables. Two cases are distinguished, based on whether the response variable is the frequency of the claims or the average cost. In both cases, however, the random forest algorithm must be applied to find the most significant variables.

3.4.1 Significant Variables for the Claims Frequency Model

First of all, it is necessary to insert a new binary variable in the dataset, called *LabelFreq*, which is 0 if the policy is without claims and 1 otherwise. After that

variable relating to the frequency of claims is created, it is better to create new variables given by the aggregation of the values of those already existing. This is because the number of tariff classes that will be subsequently created depends on the number of different values of each feature (or risk factor). Grouping the values, the number of tariff classes, on which GLM will be applied, will be reduced and therefore the computational cost of the model will also be reduced. For example, if classes were created based on *Sesso.Contraente*, which has two possible different values, and *CAP.Contraente*, which has 3811 possible different values, we would have $2 \cdot 3811$ tariff classes. Instead, if the values of variable *CAP.Contraente* were grouped in such a way as to create the variable *Zona* with values "Nord", "Centro", "Sud" and "Isole", then the number of tariff classes would be $2 \cdot 4$.

The new variables grouped are:

1. *Zona* which assumes the four values mentioned above;
2. Instead of having *Valore.Assicurato* with 4101 different values, there is *FasciaValore* with five values:
   - *1* → when *Valore.Assicurato* is between 100 and 1000;
   - *2* → when *Valore.Assicurato* is between 1001 and 2000;
   - *3* → when *Valore.Assicurato* is between 2001 and 3000;
   - *4* → when *Valore.Assicurato* is between 3001 and 4000;
   - *5* → when *Valore.Assicurato* is between 4001 and 5000.
3. Instead of having *CU.Assicurata* with eighteen different values, there is *FasciaCU* with six values:
   - *1* → it includes merit classes 1, 2 and 3;
   - *2* → it includes merit classes 4, 5 and 6;
   - *3* → it includes merit classes 7, 8 and 9;
   - *4* → it includes merit classes 10, 11 and 12;
   - *5* → it includes merit classes 13, 14 and 15;
   - *6* → it includes merit classes 16, 17 and 18.

4. Instead of having *Massimale.RCA.Sinistro* with eight different values, there is *MassimaleSin* with three values:

   - *1* → when the value of the claim maximum amount is less than 7290000€ (included);
   - *2* → when the value of the maximum amount is between 7290000€ (excluded) and 15000000€ (included);
   - *3* → when the value of the claim maximum amount is greater than 15000000€ (excluded).

   The grouping in these last three values is also done for *Massimale.RCA.Cose* and *Massimale.RCA.Persone*.

5. Instead of having *Age* with 73 different values, there is *FasciaAge* with eight values:

   - *1* → when the insured is between 18 and 21 years old;
   - *2* → when the insured is between 22 and 25 years old;
   - *3* → when the insured is between 26 and 35 years old;
   - *4* → when the insured is between 36 and 45 years old;
   - *5* → when the insured is between 46 and 55 years old;
   - *6* → when the insured is between 56 and 65 years old;
   - *7* → when the insured is between 66 and 75 years old;
   - *8* → when the insured is over 75 years old.

6. Instead of having *Marca* with 83 different values, there is *Categoria* with three values:

   - *1* → it includes all insurance policies associated with low class cars;
   - *2* → it includes all insurance policies associated with medium class cars;
   - *3* → it includes all insurance policies associated with luxury cars.

Obviously grouping values has a considerable negative side: there is a loss of information. But if this is not done, there will be a really large number of tariff classes and applying the GLMs to all of these will lead to problems.

The new variables will be used for Random Forest, but before applying it, a further step must be taken. Categorical variables without a "natural" order[5] are transformed into contingency tables using the *"dummy-cols"* command in R. For example, column *Sesso.Contraente* is transformed into two columns (because there are two values that this variable can take): *Sesso.Contraente-Femmina* and *Sesso.Contraente-Maschio*. The values that these two new variables can have are only 0 and 1, and no longer male and female as before: if in the i-th row the value of *Sesso.Contraente* is "Female" then *Sesso.Contraente-Femmina* in that row will have value 1, while the other variable will be 0.

At this point the dataset can be divided into two parts:

1. *Training Set*: is a set of data that is used to train a supervised system (such as random forest). It often consists of an input matrix with an associated response or classification. Once executed, the algorithm learns, based on the response or classification, which features discriminate the elements belonging to the different categories. Generally it includes about 70-80% of the initial dataset;

2. *Testing Set*: is another set of data, used to verify the correctness of algorithm after it has been trained on a initial training set. It is composed of observations that are not present in the training set.

To make this division, the function of R *"sample.split"* was used, which splits data from response vector Y into two sets in predefined ratio while preservig ratios of different labels in Y.

The next step is to apply the simple Random Forest as first approach. In Figure 3.10 is shown how to train a Random Forest, with the purpose of predicting the binary value of *LabelFreq* using all the other variables of *TrainingSetClaims*, and how to test the model on *TestingSetClaims*.

[5] An example is the qualitative variable *Indice.Credit.Scoring.rischio.* which has an intrinsic increasing order: 0 indicates the lowest risk, while 5 indicates the highest one.

```
RF = randomForest(LabelFreq ~ ., ntree = 100, data = TrainingSetClaims)

PredRFS = predict(RF, TestingSetClaims)
confusionMatrix(PredRFS, TestingSetClaims$LabelFreq)
```

Figure 3.10: R code: Train and test of Random Forest base with 100 trees.

A tool widely used to evaluate the quality of the classifier is the *confusion matrix*, which compares model's prediction to real solution. The structure of a confusion matrix is visible in Figure 3.11.

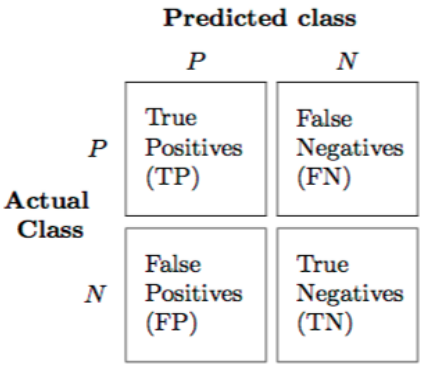

Figure 3.11: Structure of a confusion matrix.

In the case of binary response variable, assuming *LabelFreq*=0 as the positive class and *LabelFreq*=1 as the negative one, the confusion matrix is composed of four boxes, which represent different results of classification:

- *True Positives (TP)*: if the instance is positive and it is classified as positive;
- *False Negatives (FN)*: if the instance is positive and it is classified as negative;
- *False Positives (FP)*: if the instance is negative and it is classified as positive;
- *True Negatives (TN)*: if the instance is negative and it is classified as negative.

Many evaluation indexes can be calculated using these values and the most useful in this context are:

- *True Positive Rate (TPR)*: it is also called *sensitivity* and it measures the proportion of actual positives correctly identified as such, i.e. the percentage of *LabelFreq*=0 identified in the right way. It is calculated as $\frac{TP}{TP+FP}$;

- *False Positive Rate (FPR)*: it measures the proportion of negative classes wrongly identified as positive, i.e. the percentage of *LabelFreq*=1 which, however, have been classified as *LabelFreq*=0. It is calculated as $\frac{FP}{FP+TN}$.

These two indices are useful because they are used to draw the ROC (Receiver Operating Characteristic) curve, with which it is possible check the quality of the model. The model's ability to classify well is evaluated by calculating the area under the curve (Area Under Curve, AUC), whose value is between 0 and 1: 0 if the model classifies badly, 1 if the model classifies perfectly. In the case of, instead, the area is equal to 0.5, there is a random classification and the curve is a straight line called "no benefit line".

In Figure 3.12 is shown the ROC curve obtained applying standard random forest algorithm on the test dataset.

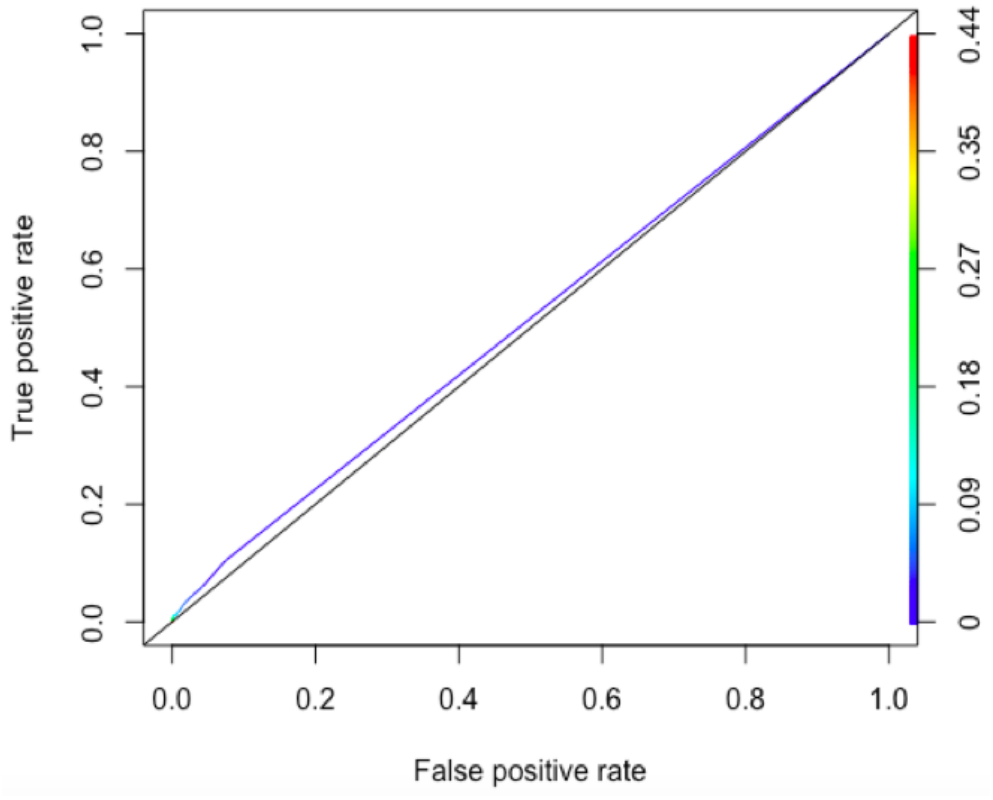

Figure 3.12: ROC curve obtained applying simple random forest on test dataset (AUC=0.515).

The classifier has a very bad performance. This is confirmed by the confusion matrix in Figure 3.13, in which it is visible that the true negatives are zero: this means that everything has been classified as non-claim. Even if the accuracy is

very high (97,19%), it is not a good classifier, because it fails to recognize when there will be a claim, which is essential for insurance. It is therefore advisable to improve the model, before seeing which are the most significant variables for it.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 31046   896
         1     0     0

               Accuracy : 0.9719
```

Figure 3.13: Confusion matrix obtained applying base random forest on test dataset.

One of the possible causes of this bad classification could be the strong unbalance of data: in fact, the dataset contains only 2,8% of *LabelFreq*=1 rows, as shown in Figure 3.14.

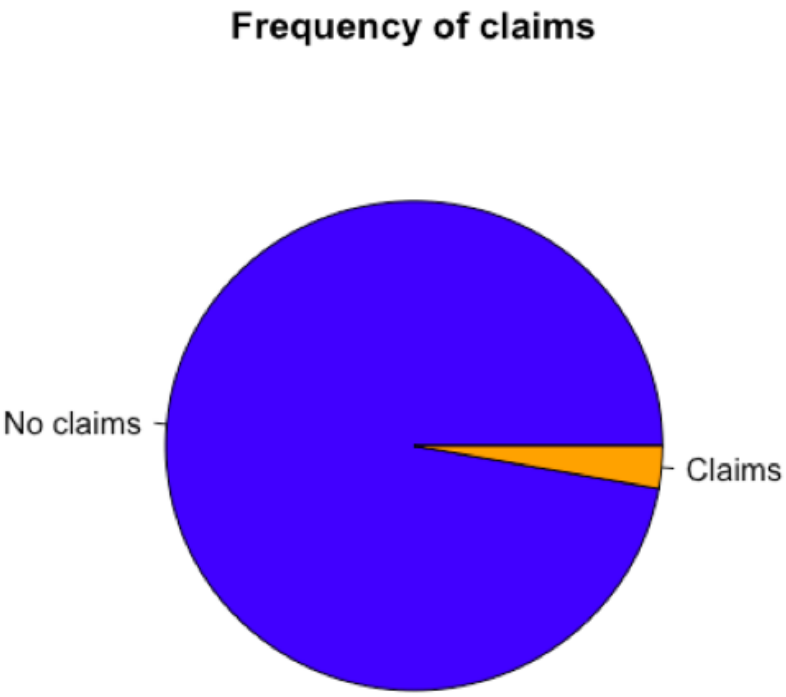

Figure 3.14: Pie chart for the frequency of claims.

It is therefore necessary to balance the dataset through downsampling technique, explained in the first part of the chapter, and reapply the random forest to the balanced dataset.

In order to get better performance, in addition to balancing the data, it is necessary to study the interactions between variables and do parameters tuning. To consider interactions, just replace $LabelFreq \sim .$ with a second degree equation $LabelFreq \sim .\hat{}2$; however, studying each interaction leads to a large increase in computational time. To improve the prediction, it would be necessary to perform tuning on all the parameters of Random Forest, but since the purpose of this model is only to select the most important variables, it is a good compromise between the computational time spent and the performance obtained by the model, doing tuning only on the *mtry* and *ntree* parameters, because they are the ones that impact the most.

In Figure 3.15 is shown how to train a Random Forest using the downsampling option provided by *caret* package in R, interactions and parameters tuning.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20274   501
         1 10816   395

               Accuracy : 0.6462
```

Figure 3.15: R code: training a Random Forest with downsampling, interactions and parameter tuning.

From the confusion matrix in Figure 3.16, it is visible that the model has improved compared to the previous one.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 22437   599
         1  8609   297

               Accuracy : 0.7117
```

Figure 3.16: Confusion matrix of Random Forest with downsampling, interactions and parameters tuning.

In fact, even if the accuracy is decreased (64,62%), now the model no longer has the true negatives equal to zero. It is also observed that the number of false positives has increased a lot, while that of false negatives has decreased slightly. This may seem like a bad thing, but it is not so: in fact, it is worse to predict a

claim as a non-claim than the opposite, because there is the risk of not having enough reserves to cover the damages. Nevertheless, care must be taken that the number of false positives is not too high.

Finally, the ROC curve and the area underlying it, observable in Figure 3.17, are better than the previous ones.

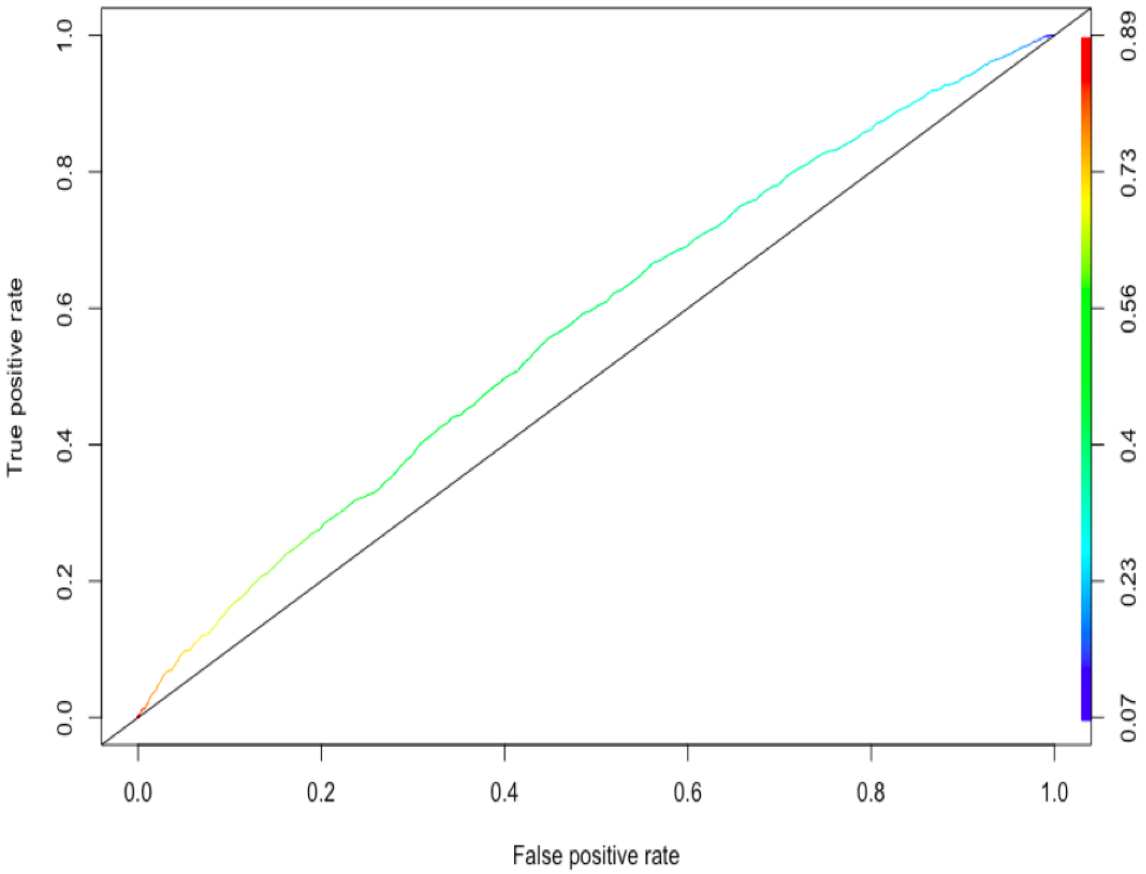

Figure 3.17: ROC curve obtained applying random forest with downsampling, interactions and tuning on test dataset (AUC=0.5726).

At this point, the most significant variables of the model are visible through the command R *varImpPlot*, in Figure 3.18.

In the plot there are not all the variables of model (considering as variables also the interactions), but only the thirty most important ones. Looking at the graph, it appears that the variables to be used for the GLM concerning the frequency of claims are: the class of merit (*FasciaCU*), the risk index (*Indice.Credit.Scoring..rischio.*), the age (*FasciaAge*) and the insured value (*FasciaValore*).

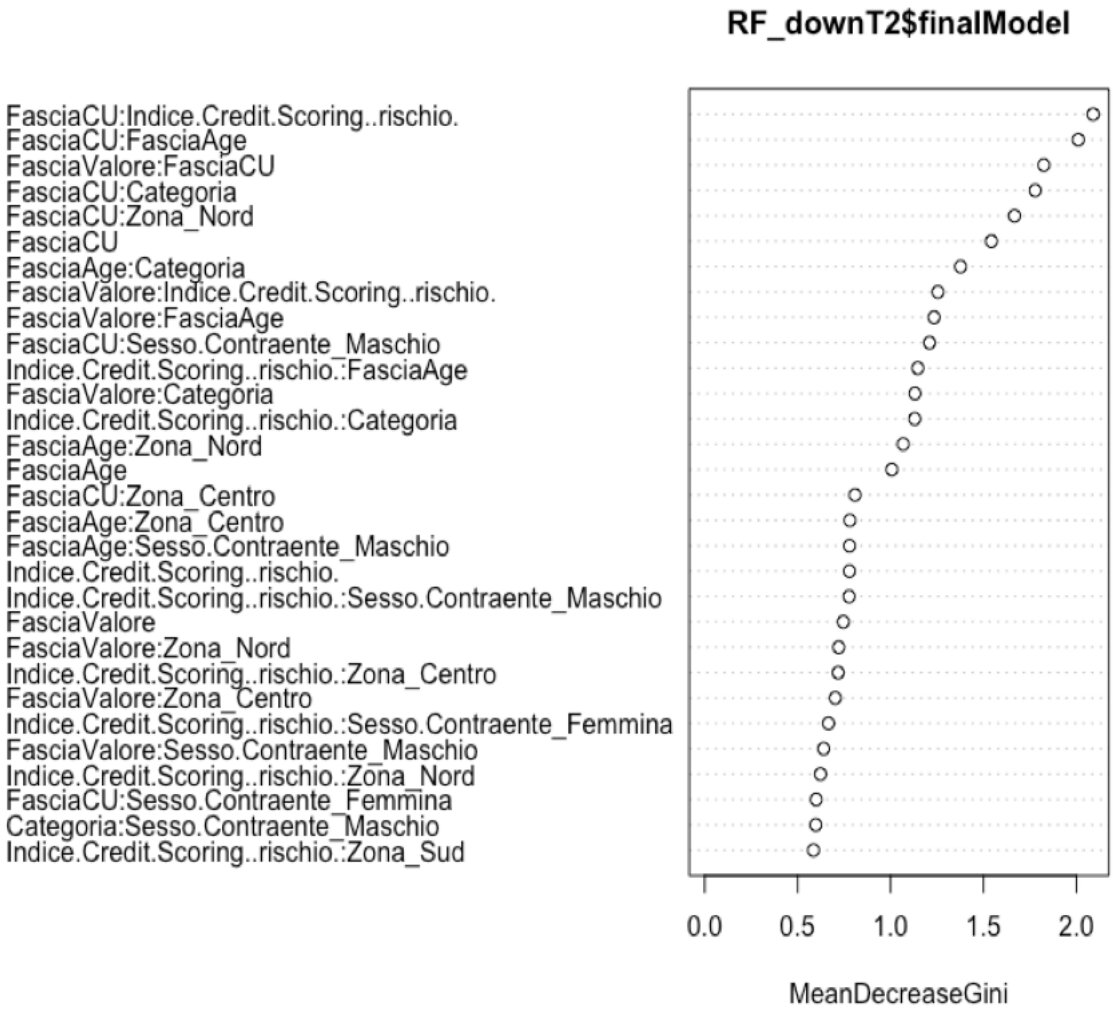

Figure 3.18: Variable importance plot of frequencies.

3.4.2 Significant Variables for the Claims Cost Model

For this model, it is necessary to take into consideration only the observations relating to policies with claims.

The basic idea is the same as before: using new variables, transforming categorical variables into contingency tables, dividing into training and testing sets and applying random forest to understand which are the most significant variables. But there are two fundamental differences:

1. this is not a problem of classification like the previous one, but of regression because the variable to predict *Sinistri.Costo.Totale* is quantitative. To evaluate the model, therefore, the confusion matrix and the ROC curve can no longer be used;

2. in this case there is no unbalancing of the data to be predicted, therefore

the downsampling technique is not used.

First, a random forest base is applied, as shown in Figure 3.19.

```
RF2 = randomForest(Sinistri.Costo.Totale ~ ., ntree = 100, data = TrainingSetDanni)
PredRFS2 = predict(RF2, TestingSetDanni, type = "response")
```

Figure 3.19: R code: Train and test of Random Forest base with 100 trees.

In the case of continuous variables, three new parameters are used to measure estimator's accuracy: the Root Mean Square Error (RMSE), the R-squared (R^2) and the Mean Absolute Error (MAE).

The first is the average of squared differences between the i-th predicted value $\hat{y}_i$ and the i-th observed value y_i, i.e.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

It tells how concentrated the data is around the line of best fit: the lower the value, the more the data follows the trend of regression line.

The R-squared, also called *coefficient of determination*, is instead an indicator that, starting from the regression line, summarizes in a single value how much the analysed magnitude deviates from this line on average. The formula is:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}$$

where $\bar{y}$ is the average of observed data. This index varies between 0 and 1: when it is 0 the model used does not explain the data at all, while when it is 1 the model perfectly explains the data.

Finally, MAE is the average over test sample of all absolute errors, which are the absolute values of the differences between predictions and actual observations. The formula is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Like RMSE, it is a negatively-oriented score, which means that lower values are better.

The three indices relating to the random forest base model are:

```
> RMSE
[1] 25031.05
> RSquared
[1] 2.275e+32
> MAE
[1] 1477.963
```

Figure 3.20: R code: Root Mean Square Error, R-squared and Mean Absolute Error for Random Forest base model.

These values suggest that the quality of the model is not very high and that it is necessary to improve it for better results. To do this, the same method used for the claims frequency model is used: tuning on the *mtry* parameter is done and interactions are considered. The result is the following:

```
> RMSE_TInt
[1] 24245.45
> RSquared_TInt
[1] 2.134439e+32
> MAE_TInt
[1] 1374.904
```

Figure 3.21: R code: Root Mean Square Error, R-squared and Mean Absolute Error for Random Forest with tuning and interactions.

Looking at the new index values, it is stated that the results are slightly better than before, but the fact remains that the overall result of the model is not the best. It should be remembered, however, that the purpose of the model is not to predict, but to understand which are the characteristics that have the greatest impact on the cost of the claim. For this reason, no attempt is made to further improve the model and the *varplot* is observed directly.

From Figure 3.22 it can be concluded that the most significant variables for the cost model are the insured value (*FasciaValore*), the car brand (*Categoria*), the class of merit (*FasciaCU*) and the risk index (*Indice.Credit.Scoring..rischio.*). Note that they are not the same as those found for the frequency model.

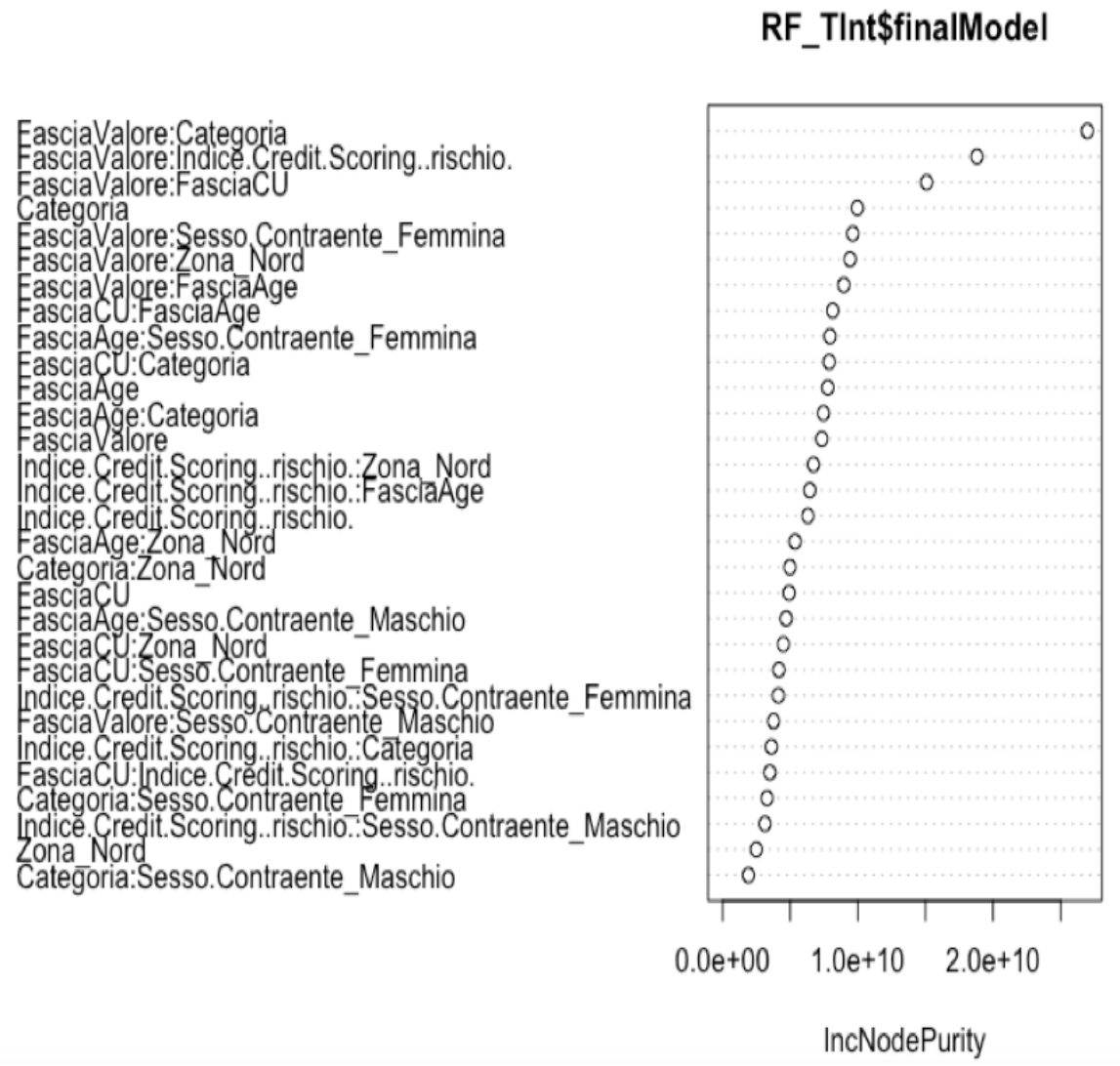

Figure 3.22: Variable importance plot of costs.

3.5 Application of GLMs

3.5.1 Creation of Tariff Classes

As mentioned in the first chapter, the insurer divides the community of risks into subgroups or tariff classes, which have similar characteristics, so as to be able to attribute the same technique to the risks belonging to the same class. Through this process the premiums are therefore differentiated for the insured, depending on the different risk profile.

In this thesis it was decided to divide the portfolio according to five risk factors (selected on the basis of the results of the two previous Random Forest): age, brand, class of merit, insured value and risk index. As said previously, these

risk factors are all classification variables divided into levels; in fact, *FasciaAge* has eight levels, *Categoria* has three, *FasciaCU* has six, *FasciaValore* has five and *Indice.Credit.Scoring..rischio.* has six. The portfolio is therefore made up of 4320 tariff classes ($8 \cdot 3 \cdot 6 \cdot 5 \cdot 6$), identified by the determinations of the five tariff classification variables. It should be noted that 4320 is the number of tariff classes that should be there, but it is not necessarily the actual number. In fact, it could be that there are no policies with the characteristics of a certain tariff class and so, instead of leaving the line corresponding to that class empty, it was decided to not insert it into the new dataset. In this case, instead of having 4320 rows, the dataset used for the GLMs has 2735.

For each tariff class, the exposure, the number of claims observed, the total damage, the number of policies and the claims frequency must then be indicated. Figure 3.23 shows an extract of the new database (*DatasetGLM*) composed of tariff classes and related information.

FasciaCU	FasciaAge	FasciaValore	Rischio	Categoria	EspTot	NumClaims	NumPolizze	DannoTot	FreqOsservata	CostOsservato
1	5	1	0	1	720.589041	19	772	33979.00	0.02636732	1788.368
1	5	1	0	2	204.413699	7	216	16220.80	0.03424428	2317.257
1	5	1	0	3	5.326027	1	6	2285.00	0.18775720	2285.000
1	5	1	1	1	674.564384	16	730	30065.14	0.02371901	1879.071
1	5	1	1	2	219.991781	4	232	16522.00	0.01818250	4130.500
1	5	1	1	3	11.372603	0	13	0.00	0.00000000	NaN

Figure 3.23: Extract from the new dataset *DatasetGLM*.

For each line, thanks to the application of GLM, the expected number of claims and the expected value of compensation for individual claims are estimated. Through the first estimated value the frequency of claims is found and this is multiplied with the second estimated value to determine the fair premium within each class.

Let us remember that in GLMs the response variables Y_i are supposed to be stochastically independent, with distributions belonging to the same exponential family. The expected value of Y_i is linked to the determinations of the explanatory variables $\mathbf{x}_i'$ by a invertible link function g.

Therefore:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i$$

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\eta_i)$$

$$i = 1, ..., n$$

where $\eta_i = \beta_0 + x_{i1}\beta_1 + ... + x_{im}\beta_m$ is the linear predictor, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)'$ is the vector of parameters that will be estimated with the regression and $\mathbf{x}_i' = (1, x_{i1}, ..., x_{im})$ is the vector of determinations of the explanatory variables. Let us also remember that the link function that allows to reach the construction of a multiplicative model, which is the one used in the insurance world, is the logarithm. Moreover, the logarithm allows to obtain always a positive expected value.

Now, based on what has just been described, it is possible to proceed with the construction of the two GLMs.

3.5.2 GLM for the Number of Claims

The GLM used to estimate the number of claims in each tariff class k coincides with the model described in the Paragraph 2.2.4. Therefore, the number of expected claims for each tariff class will be estimated using a model in which the response variables M_k, i.e. the random number of claims that will affect the insured risk, are considered stochastically independent and with Poisson distribution.

The expected value of response variables is $\mu_k = \mathbb{E}[M_k] = t_k\lambda_k$, that is to say that the expected value of the number of claims in class k corresponds to the product between the total exposure of the class t_k and the expected annual number of claims λ_k for each insured in one year.

Linear predictors of tariff classes are indicated by $\eta_k = \ln t_k + \mathbf{x}_k'\boldsymbol{\beta}$, in which $\ln t_k$ is the *offset* term, introduced in order to take into account the exposure since in the data there are not only unit exposures, but also exposures lower or higher than a year (as seen in Figure 3.24).

As link function, the canonical one for the Poisson distribution is chosen, i.e. $g = ln$.

After dividing the dataset into training and testing set, the model in R is estimated through *glm* command and the code is visible in the Figure 3.25.

In the command, *model.freq* is the name assigned to the regression output, *NumSin* is the response variable, i.e. the number of claims, estimated as a function of tariff variables. The logarithm of exposure (offset term), is also added to

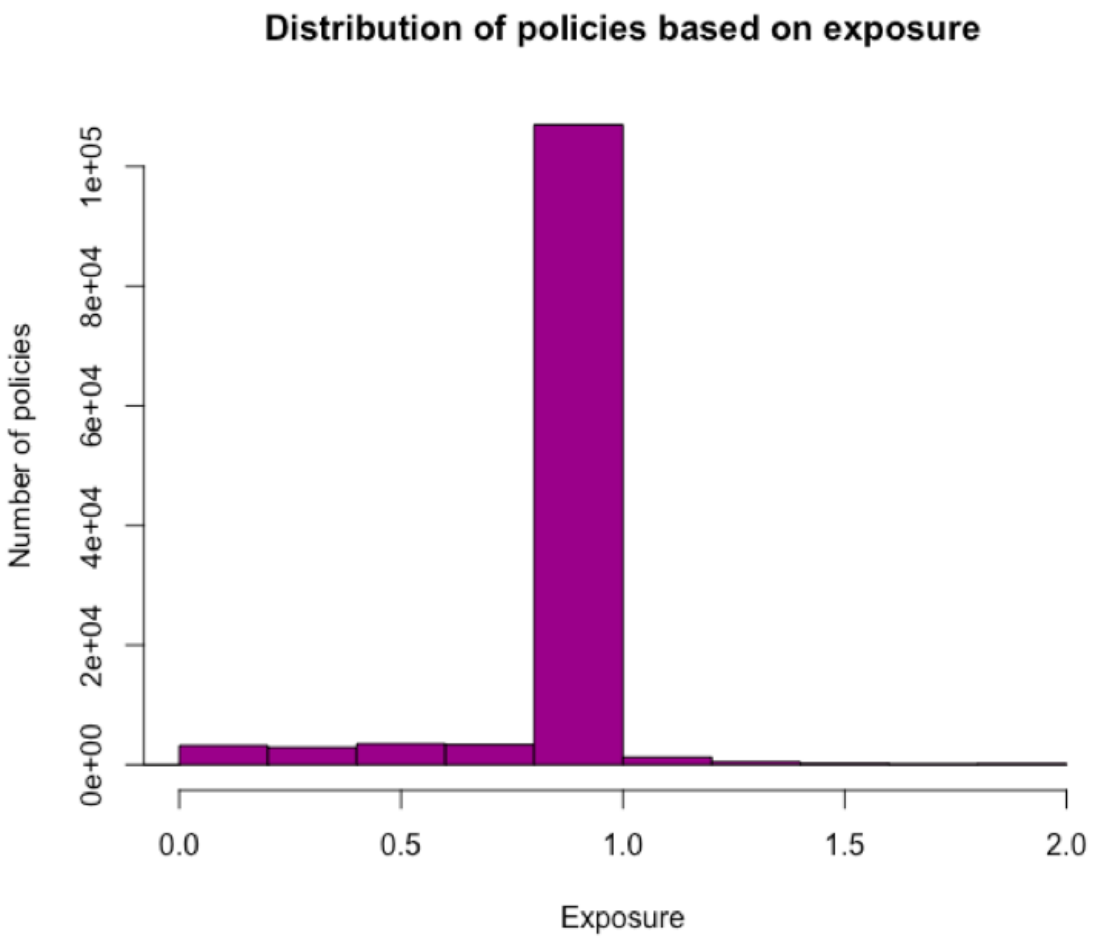

Figure 3.24: Histogram of the number of policies based on exposure.

```
model.freq <- glm(NumClaims ~ FasciaAge + FasciaCU + FasciaValore + Rischio + Categoria +
                  offset(log(EspTot)), data = TrainSetGLM, family = "poisson")
```

Figure 3.25: R Code: Generalized Linear Model for the frequency of claims.

the explanatory variables.

Finally, the distribution assigned to the response variables is specified with *family = poisson*. It is not necessary to specify the logarithmic link function, since the program automatically recognizes the logarithm as a canonical link.

Once the model has been trained on *TrainSetGLM*, it is tested on *TestSetGLM*. The results obtained from the prediction, using the *predict* function, are however the linear predictors; to get the number of claims, just consider the exponential of these.

To evaluate the quality of the model, RMSE and MAE are considered, as done previously. In this case, they are respectively 0.1332534 and 0.6322603 and, being low values, it can be said that the model predicts well.

As a last step, it is necessary to calculate the frequency of claims, dividing the number of claims estimated by the total exposure. Once the frequency is found, proceed by calculating the average cost per claim.

3.5.3 GLM for the Average Cost per Claim

For the estimation of the average damage per claim in the tariff classes, instead, the GLM will have the structure envisaged by the model explained in the Paragraph 2.3.4. Also in this case, the response variables Y_k, which indicate the average damage per claim for the different tariff classes, are supposedly stochastically independent. But now they will have a Gamma distribution and the weight introduced for their estimate will be m_k, that is the total number of claims that hit class k.

In this model the expected value of the average damage per claim will be given by $\mathbb{E}[Y_k] = \mu_k = g^{-1}(\eta_k)$, with $g = ln$. Actually the canonical link function for the Gamma distribution is $g = -\frac{1}{\mu}$, but the latter requires placing constraints on the regression parameters to make the linear predictor negative and the expected value positive. It is therefore preferable to set a logarithmic link function that allows to obtain positive values for the expected value, without having to place constraints on the parameters. Furthermore, as stated above, the choice of the logarithmic link function depends on the fact that, in insurance practice, tariff models of a multiplicative type are often adopted and the logarithm allows the construction of this type of model.

In R the model is estimated using the command shown in the Figure 3.26.

```
model.cost <- glm(CostOsservato ~ FasciaAge + FasciaCU + FasciaValore + Rischio + Categoria +
      offset(log(EspTot)), data = TrainSetGLMC, family = Gamma(link = "log"), weights = NumClaims)
```

Figure 3.26: R Code: Generalized Linear Model for the average cost.

Similarly to what was seen for the model for the number of claims, *model.cost* indicates the name assigned to the regression output and *CostOsservato* is the estimated response variable depending on the tariff variables. With *TrainSetGLMC* the dataset from which to obtain the data to perform the procedure is indicated; for the estimation of the model, tariff classes in which compensation has been paid by the insurance company will be considered, since the constraint

of positive compensation has been imposed (*TrainSetGLM$CostOsservato > 0*).

Finally, the distribution assigned to the response variables is specified with the command *family = Gamma(link = "log")* and it is also specified that the logarithmic link function should be applied to the model, and not the canonical connection for the Gamma distribution.

Also in this case the GLM is subsequently applied on the testing set and the quality is evaluated through the RMSE and MAE, which are respectively 5747.048 and 2475.822. The quality of prediction is not bad, but it is not as good as the previous one. This is because it is very difficult to predict the exact amount of the damage. In fact from the results of the model, it emerges that the damage depends more than anything from the category of the vehicle: from this it obviously follows that the damage amounts will be greater if luxury cars are damaged. For example, if there is an accident in which a mirror of a Grande Punto Fiat is broken, it is known that it will be paid less than to make an accident in which the mirror of a Ferrari is broken. The problem, however, is that the cost varies greatly depending on the type of accident done. In fact, it can not be known if the driver of the Ferrari will make an accident in which he will only break the mirror or one in which he will destroy the car, but according to this the cost obviously changes a lot. For this reason, therefore, it is difficult to accurately predict the cost of a claim. To improve the forecast it should have information about the type of accident that will happen, but these are impossible to have.

3.6 Combination of the Two Models for the Fair Premium

To calculate the fair premium for each tariff class are necessary the estimated claim frequencies and the average damage amount per claim. An extract of these values, together with those observed, are visible in Figure 3.27. It should be noted that the values that have been chosen to show are the results of the models applied on the testing set, even if they are not the best. In fact, the results of the training set model are better than those on the testing set, because the model is evaluated on the same data on which it was trained. But in a real case, the model

is evaluated on "new" data, with which it has never interacted, as happens for the testing set.

	FasciaCU	FasciaAge	FasciaValore	Rischio	Categoria	FreqOsservata	CostOsservato	FreqPredetta	CostPredetto
1	1	5	1	0	1	0.02636732	1788.368	0.02608728	1177.9662
2	1	5	1	2	2	0.04931221	1805.666	0.03504498	1615.7859
3	1	5	1	4	1	0.02654932	1930.352	0.02577780	707.0163
4	1	5	4	0	2	0.03743809	1493.874	0.03769786	3274.9452
5	1	5	3	1	1	0.02592581	1523.505	0.02719773	1190.3330
6	1	5	3	2	1	0.02842679	1337.167	0.02711671	957.1807

Figure 3.27: Comparison between observed and predicted frequency and cost of testing set.

The last step to calculate the fair premium for each tariff calss is to multiply the predicted claim frequencies by the average cost per claim. In Figure 3.27 the result of multiplication is visible.

	FasciaCU	FasciaAge	FasciaValore	Rischio	Categoria	PremOsservato	PremPredetto
1	1	5	1	0	1	47.15448	30.72994
2	1	5	1	2	2	89.04140	56.62518
3	1	5	1	4	1	51.24951	18.22533
4	1	5	4	0	2	55.92780	123.45842
5	1	5	3	1	1	39.49808	32.37436
6	1	5	3	2	1	38.01136	25.95559

Figure 3.28: Calculation of the fair premium on the testing set and comparison between observed and predicted premium.

The predicted premium is quite similar to that observed, so the model is quite good; this is confirmed by the RMSE and the MAE, which are respectively 2907.752 and 410.5310. To improve the final result, the prediction of the average cost should be improved.

Lastly, remember that the result obtained is not the final premium of an insurance policy, but the fair premium. To this must then be added earnings, fixed management costs and taxes to obtain the final premium.

In the next chapter it will be explained in a theoretical way how to find the earnings of an insurance company and therefore how to optimize the premium in such a way that the earnings of the insurance company are as high as possible.

Chapter 4

Pricing Strategies for Insurance

As mentioned in the other chapters, once the reserves that insurance company has to set aside to cover accident costs have been estimated, i.e. the fair premium, insurers add a load at this price to make profits and cover their expenses. The insurance company must define a pricing strategy in order to choose a loading that optimizes its wealth, taking into account different factors.

In general, the pricing strategy is a way to find a competitive price for a product or service, in this case an insurance, which results in the increase of company's profit margins and volume of business and also in the maximization of customer's lifetime value and loyalty.

There can be various types of pricing strategies:

- The simplest case is one in which neither consumers nor competitors are taken into account, but only costs. The production cost of an item is evaluated and to this is added a *mark up*, that is a margin, to determine the selling price.

- An evolution of the previous strategy is that which also takes into account consumers. In this case the price of the product derives from the equilibrium of *supply* and *demand*, which occurs when the quantity demanded of an item equals the quantity offered of the same item.

  In microeconomics, demand is the amount of consumption required by the market and by consumers of a product or service, given a certain price. This is influenced by several factors, first of all the price of the purchased

item. Other influencing factors are the price of complementary products, the consumer's income and his needs.

The offer, on the other hand, is the quantity of a product or service that is put up for sale at a given time and at a certain price. This is influenced by prices, production costs, technology and government policies.

It should be noted that in the event of a monopoly, that is when there is only one seller, this case leads back to the simplest case.

- The most complete strategy is one that also takes into account the role of competition. In the classical microeconomics literature game theory[1] is used to model the interaction between two competitors.

In this chapter the last two types of strategies will apply to the field of insurance. The first pricing strategy is not considered because it is too simplistic.

However, before explaining the models, it is well defined what the demand elasticity is, on which practically all pricing strategies are based.

The *price-demand elasticity* ϵ measures the variation in the quantity demanded of a economic item in reaction to a change in its price. This value is always indicated in absolute value and it can take values between zero and infinity. Based on the value, various types of demand can be distinguished:

- if $\epsilon = 1$, there is a *demand for unitary elasticity* and the demand is exactly proportional to the price changes;
- if $\epsilon < 1$, there is an *inelastic demand* and the quantity demanded is little influenced by price changes;
- if $\epsilon > 1$, there is an *elastic demand* and the quantity demanded is influenced by price changes in a more than proportional ratio;
- if $\epsilon = 0$, there is a *rigid demand* and the quantity demanded does not vary with the price, i.e. whatever the price, the consumer always buys the same quantity of the item. A demand of this kind can be had in the

[1] Game theory is a discipline of applied mathematics that studies and analyzes the individual decisions of a subject in situations of conflict or strategic interaction with other rivals aimed at the maximum gain of each subject.

case of mandatory insurance, like the Car Liability, and monopoly. This case, however, is not realistic, because in reality there is competition and therefore the quantity demanded changes a lot based on the price, even if it is a mandatory insurance;

- if $\epsilon = \infty$, there is a *perfectly elastic demand* and the demand is hypersensitive to price changes, i.e. a small price change affects the consumer's purchase decision.

4.1 Pricing Strategy for Non-Life Products Based only on Customers

The sensitivity of insurance customers to the price and the way in which this affects its variation has been subject of an extensive analysis in the research literature of the insurance market. Several methods have been developed to understand which customers are most valuable and in this thesis it was decided to describe the method that uses statistical tools to estimate the potential value of insurance customers and their price demand elasticity.

4.1.1 Estimation of the Potential Value of the Customer

The first step of this method is to segment the available customer portfolio based on the potential value of the customer. In this way the insurance company understands which customers are potentially valuable, and therefore on which it is better to invest, and which are the non-valuable customers on which it must minimize investments.

The potential value is a measure defined as the sum of the current and the future value of the customer. The current value is given by the difference between two elements: the sum of the premiums paid during the entire relationship with the insurance company for all insurance products and the sum between the claims that occurred in the same period and the acquisition and issue costs of these policies. The future value of the customer, on the other hand, is the expected value of the margin that the customer will leave in the future relationship with the

company, i.e. the premiums minus expected losses, expenses and commissions. The formula to find the future value for k insurance products, considering that cross-selling[2] is possible, is

$$Future\ Value = \sum_{j=1}^{k}(p_j margin_j) \sum_{t=1}^{T} \left(\frac{1}{1+r}\right)^t \left(\frac{l_t}{l_0}\right)$$

where p_j is the probability that the customer will continue to buy the j product with the current insurance company and $margin_j$ is the margin for that product. In the second summation, instead, the first term represents the discount interest rate for a certain year t, while the second term is the expected loss ratio of purchased insurance products.

The information needed to calculate a customer's potential value is all in the possession of the insurance company, except the likelihood that the customer will continue to buy the product in the future. But to see what this probability is, it is just a different way to see what is the market elasticity. So to calculate it, it is necessary the demand function, which corresponds to a GAM-Logit model. It is a logistic regression that uses the Generalized Additive Models to predict the probabilities (according to relevant explanatory variables), but it will not be explained in detail since this is not the main objective of this thesis.

4.1.2 Segmentation of Customer Portfolio

Once the potential value of each customer has been calculated, the portfolio is segmented. There is no single correct way for the division of the insured, but in this thesis it was decided to segment into four groups according Figure 4.1.

The strategies to be adopted based on segmentation are briefly discussed.

- *Segment I*: it can be considered unattractive considering that it has low potential value and low current value. Future profitability is expected to be low and in order to improve it, strategies should focus on cost reduction and possibly on less promotions, instead of trying to increase the purchases;

[2]It is a sales strategy consisting in proposing to the customer who has already purchased a particular product or service also the purchase of other complementary items.

		CURRENT VALUE	
		Low	*High*
CUSTOMER POTENTIAL VALUE	*High*	II	IV
	Low	I	III

Figure 4.1: Segmentation with current value and customer potential value.

- *Segment II*: it has high potential value, but low current value. Companies should aim to get a larger part of the customer's potential in this segment, for example by doing up-selling[3]. In this way the company will increase the customer's profitability by increasing its share of purchases;

- *Segment III*: it has low potential value and high current value. In this case, the company deals with relatively loyal customers and the likelihood of successfully applying up-sell is low. Since customer loyalty is very important to companies, they must try to keep these customers;

- *Segment IV*: it contains the most valuable customers, who are loyal and have a high potential value. Not having this group of customers anymore would bring a serious loss to the insurance company, for this reason the management must strive to maintain these customers (perhaps giving them some benefits that customers of other segmentations do not have).

4.1.3 Tariff Optimization Model

After splitting the insured into the four segments, it is possible to apply an algorithm that optimizes the premium for each customer group. The premium is defined as optimal if it maximizes the margin, subject to two constraints:

[3]It is a technique which consists in proposing to the customer superior qualitative versions of the product or service initially requested

1. The Customer Retention Rate (CRR) should be maintained above a predefined value. It expresses the percentage of customers that continues to buy in a given time interval. More in detail, taking a given time frame as a reference, the CRR is determined by the percentage ratio between the number of customers in the portfolio at the end of the period examined and the number of the customers present in the portfolio at the beginning of the same period;

2. The optimal increment for each group of customers must be kept within a predefined range; the extremes of the range vary according to the group considered and are decided by the management.

The optimization problem can therefore be written as follows:

$$max\ \pi_i = f(r_i, d_i)$$

$$s.a.\ \ r_i \geq \bar{r}$$

$$d_{min} \leq d_i \leq d_{max}$$

where π_i is the objective function equal to $premium_{t,i}(1 + d_i) - \mathbb{E}[losses] - \mathbb{E}[expenses\ and\ commissions]$ and it represents the margin. It should be noted that $premium_{t,i}$ is the actual net premium and that $\mathbb{E}[losses]$ is the fair premium estimated in the previous chapter. Furthermore r is the CRR, d represents the optimal percentage increase of the premium, while $\bar{r}$, d_{min} and d_{max} are the values pre-established by management.

The optimization algorithm starts by generating random numbers, which represent the premium increase, for each customer segment. These numbers are generated by a distribution chosen by the user, which is a uniform U[d_{min}, d_{max}] by default. For each d_i, the renewal probability p_i is considered: if this is lower than the one estimated at the end of Section 4.1.1 then the i-th increase rate is excluded from the model, otherwise, if the probability renewal is equal or higher, the increase rate is taken into account. For each of the values not discarded, it is then necessary to verify the CRR: if, when the optimal increase is applied, the retention rate is less than the predefined value $\bar{r}$, this increment is rejected. Finally, among the increase rates that were not discarded in the last step, the one that makes the margin higher is chosen.

This procedure is repeated n times (each time different random numbers are generated). The optimal renewal premium is obtained by the multiplication between the current premium and the optimal increase rate found, such that it generates the highest margin value of all the iterations of the algorithm.

This model, however, does not take into account a very important factor for pricing strategies: the competition. Generally, if a company X increases its rate level, the quantity of policies sold decreases, but if the current rate of X is lower than that of competitor Y, an increase in rate level (provided that, even with the increase, it remains lower than the competitor) may not lead to a decrease of sales.

So the prices of competitors play an important role and for this reason the model just seen is considered incomplete in reality and a model that also takes into account the competition of the insurance company has been developed.

4.2 Pricing Strategy for Non-Life Products in Competitive Markets

In recent years the competition between insurance companies has increased more and more and if a company wants to survive, it must be able to define the prizes in order to effectively respond to the premiums offered by competitors. For this reason, it is necessary not only calculate the optimal premium strategy for an insurance company (as was done before), but also ask how this strategy is related to the competitive insurance market.

Before describing the various models that have been developed for a pricing strategy in a competitive market, a glossary of the symbols that will be used is drawn up.

Glossary

$\{V_k\}_{k\in\mathbb{N}}$: is the sequence of the *volume of business* underwritten by the insurer in year [k, k+1). It can be measured in any significant unit.

$\{\pi_k\}_{k\in\mathbb{N}}$: is the sequence of the *break-even premium* in year [k, k+1), i.e. risk

premium (or fair premium) plus expenses per unit exposure.

$\{p_k\}_{k\in\mathbb{N}}$: is the sequence of the *premium* charged by the insurer in year [k, k+1). It is the decision-making parameter.

$\{\bar{p}_k\}_{k\in\mathbb{N}}$: is the sequence of the *"average" premium charged by the market* in year [k, k+1). It is assumed that this process is stochastic.

$\{w_k\}_{k\in\mathbb{N}}$: is the sequence of the *company's wealth* in year [k, k+1).

$\{\gamma_k\}_{k\in\mathbb{N}}$: is the sequence of the *reputation's impact* on the volume of business in year [k, k+1) and $sign(\gamma_k)$ is the sign of this parameter which represents the kind of impact that reputation has on the volume of business.

$\{\theta_k\}_{k\in\mathbb{N}}$: is the sequences of the *set of all other stochastic variables* (as inflation, interest rate, marketing etc), that are assumed Gaussian and independently distributed in time, and that are considered relevant to the demand function in year [k, k+1).

$\{a_k\}_{k\in\mathbb{N}}$: is the sequence of the *excess return of capital* in year [k, k+1).

r: is the *rate of return* on equity required by shareholders of the insurer whose strategy is under consideration. It is assumed that this rate is deterministic.

v: is the *corresponding discount factor* and it is equal to $v = (1 + r)^{-1}$.

4.2.1 Taylor's Model

In the literature, various models concerning the competition in the insurance market have been discussed. The first was that of *Taylor G. C.* in 1986, which explored the relationship between the behavior of the Australian market and the optimal response of an individual insurer, whose purpose is to maximize the wealth of the company. He has assumed that this relationship depends on different factors, including:

1. The predicted time which will elapse before having profits;

2. The price elasticity of demand for the insurance product under consideration. Moreover, he has assumed that the policies display a positive price-elasticity of demand. This means that, if the market mainly begins underwriting at a loss, any attempt to maintain profitability of the company will result in a reduction of his volume of business;

3. The rate of return r required on the capital supporting the insurance operation. In general, r is the net gain or loss on an investment over a specified time period and it is expressed as a percentage of the investment's initial cost.

Consequently, for a given sequence of average market prices, the demand function $f_k(.)$ is given by a relation of the following type

$$V_k = f_k(p_k, \bar{p}_k, V_{k-1}, \theta)$$

It is visible that the demand is affected by the price p_k of policies, from the average price offered by the competitors $\bar{p}_k$ and other factors θ such as the law for certain coverages (liability insurance on cars is legally required in all states) or the customer's willingness and ability to pay.

Moreover, the objective function to be maximized is the expected present value of the wealth of company in a pre-defined finite time horizon.

$$\sum_{k=1}^{K} v^{k-\frac{1}{2}} V_k(p_k - \pi_k)$$

where the various variable are those described in the glossary. The previous demand function, however, is too general to get useful results, so Taylor restricted it to

$$V_k = f_k(p_k, \bar{p}_k, V_{k-1}, \theta) = V_{k-1} f(p_k, \bar{p}_k)$$

In this restriction there are implicit assumptions:

- the demand function is stationary over time, for this reason the subscript in f_k has been dropped;
- V_k is assumed to be proportional to V_{k-1};
- the discarding of the unspecified set of variables is equivalent to treating $\bar{p}_k$ exogenous to the strategy of the insurer under consideration.

After these restrictions the objective function becomes:

$$\sum_{k=1}^{K} v^k (p_k - \pi_k) \left[\prod_{j=1}^{k} f(p_j, \bar{p}_j) \right]$$

so the optimal pricing strategy prescribes a sequence of premiums p_k such as to maximize the expected profit discounted at rate of return per annum.

The results he found are very interesting. Firstly, he showed that as v increases, the project future premium involved in the optimal underwriting strategy also increases. Moreover, if this v is sufficiently large, the optimal strategy will never involve any loss cases (when the premium is lower than the break-even premium).

According to the results of his studies, he concluded that the optimal strategies do not follow what someone might expect. In fact, contrary to what can be expected, during a period of depression of the premium rates it is not said that there is no profitability for an insurer. For example, if it is considered the case in which the average market rate is below the break-even point (and therefore competitors sell a policy at a price lower than its cost), it may be thought that the best choice of an individual insurer is to sell off the policy too, but in reality it is not certain that this is so. So his idea is based on the fact that to earn more profits an insurance company must study the behavior of the market, but it does not always have to follow it.

Taylor's idea was later extended in 2015 by Pantelous and Passalidou, who developed a model that takes into account both the competitive market and the company's reputation.

4.2.2 Model of Pantelous and Passalidou

With this approach the volume of business has a discrete-time stochastic demand function (because θ_k and $\bar{p}_k$ can be considered as random variables), like in the previous model, but the formulation of the business volume uses also the reputation of the company. This is a factor that should not be excluded, because it makes the model more realistic: in fact, the reputation of the insurance company has a strong influence on the customer's purchase decisions or, in other words, on the company's product demand.

Firstly, according to Taylor's model, the following three assumptions are made:

1. There is positive price-elasticity of demand, i.e. if the market begins underwriting at a loss, any attempt by a particular insurer to maintain profitability will result in a reduction of his volume of business;

2. There is a finite time horizon T;

3. The demand of the year $k+1$ is proportional to the demand of the previous year k.

Secondly, the purpose of the model is to determine the strategy that maximizes the expected total utility of the wealth at time k over a finite time horizon T, i.e.

$$\max_{p_k} \ \mathbb{E}[\sum_{k=0}^{T} U(w_k, k)]$$

where $U(w_k, k) = v^k w_k$ is the actual value of the wealth.

Very important is the wealth process that is

$$w_{k+1} = -a_k w_k + (p_k - \pi_k) V_k \tag{4.1}$$

where a_k represents the excess return on capital, i.e. the return on capital required by the shareholders of the insurer. So the wealth of the company is given by the profit on the policies less the interests that the company has to pay to the shareholders. Moreover, the volume of business V_k is given by

$$V_k = V_{k-1}\left(\frac{\bar{p}_k}{p_k}\right)^{\alpha} + sign(\gamma_k)|\gamma_k|^{\beta} e^{\theta_k} \tag{4.2}$$

The first part of the sum in (4.2) concerns the competitive market and it is the same as that of the previous model, with the exception of the parameter α. This sensitivity parameter models the elasticity of the company's volume of business to a change in premium in the preceding year k. The second part of the sum, instead, incorporates the parameter γ_k and the sensitivity parameter β that models the effects of reputation on the volume of business: sign of γ_k is equal to +1 when the company has a good fame or it is equal to -1 when the reputation is bad. However, the volume of business is exponentially affected by the stochastic parameter θ_k, which comprises all variables that are relevant to the demand function in year [k, k+1) and represents the white noise.

Having said that, one can derive the optimal premium through the following Theorem (of Pantelous and Passalidou), whose demonstration is omitted.

Theorem. For the wealth process given by

$$w_{k+1} = -a_k w_k + (p_k - \pi_k)\left(V_{k-1}(\frac{\bar{p}_k}{p_k})^{\alpha} + sign(\gamma_k)|\gamma_k|^{\beta}\right) \tag{4.3}$$

and for the maximization problem defined by

$$\max_{p_k} \ \mathbb{E}[\sum_{i=k}^{T-1} v^i w_i] \tag{4.4}$$

with initial conditions w_0, V_0, V_{-1}, a_0 and γ_0, the optimal premium $p*_{max,k}$ is given as a solution to the polynomial when

(a) $\alpha > 1$,

$$p^{\alpha+1}{}_k + b_1 p_k + b_2 = 0 \ \textit{and} \ 0 < p_k < \left(1 + \frac{2}{\alpha - 1}\right)\pi_k, \ \ \textit{for } k = 0, 1, ..., T-1 \tag{4.5}$$

It is interesting to note that the optimal premium has an upper bound that is related to the break-even premium π_k and the elasticity parameter α.

(b) $0 < \alpha \leq 1$,

$$p^{\alpha+1}{}_k + b_1 p_k + b_2 = 0 \ \ \textit{for } k = 0, 1, ..., T-1 \tag{4.6}$$

In this case there is always an optimal premium.

The parameters b_1 and b_2 are $\frac{(1-\alpha)V_{k-1}\mathbb{E}[\bar{p}_k^{\alpha}]}{sign(\gamma_k)|\gamma_k|^{\beta}\mathbb{E}[e^{\theta_k}]}$ and $\frac{\alpha\pi_k V_{k-1}\mathbb{E}[\bar{p}_k^{\alpha}]}{sign(\gamma_k)|\gamma_k|^{\beta}\mathbb{E}[e^{\theta_k}]}$, respectively, while γ_k is given by $w_0 d_0 + e_0$. Moreover, it is defined d_k as $v^k - a_k d_{k+1} > 0$ with $d_T = 0$ and e_k as $(p*_{max,k} - \pi_k)\left[V_{k-1}\frac{\mathbb{E}[\bar{p}_k^{\alpha}]}{(p*_{max,k})^{\alpha}} + sign(\gamma_k)|\gamma_k|^{\beta}\mathbb{E}[e^{\theta_k}]\right]d_{k+1} + e_{k+1}$ with $e_T = 0$.

The root of the polynomial given by (4.5) and (4.6) is the optimal premium and it must be a real and positive number. For this reason, it is necessary to consider only the cases in which the result is positive and real.

The first case to consider is where $sign(\gamma_k) = 1$. This has two possible sub-case based on the value of the polynomial's coefficients:

- For $\alpha > 1, b_1 < 0$ and $b_2 > 0$, the polynomial has zero or two positive roots;
- For $0 < \alpha \leq 1, b_1 > 0$ and $b_2 > 0$, the polynomial has no positive root. So this case must not be taken into consideration.

The second case to consider, instead, is where $sign(\gamma_k) = -1$. Also in this case there are two possible sub-cases:

- For $\alpha > 1, b_1 > 0$ and $b_2 < 0$, the polynomial has exactly one positive root;
- For $0 < \alpha \leq 1, b_1 < 0$ and $b_2 < 0$, the polynomial has exactly one positive root.

Combining the Theorem with the request to have a positive root, the following corollary is obtained.

Corollary. For the parameters of the Theorem, there is an optimal solution $p*_{max,k}$:

I For $sign(\gamma_k) = 1$, with $\alpha > 1$ and $0 < p_k < \left(1 + \frac{2}{\alpha-1}\right)\pi_k$, when the polynomial (4.5) has two positive roots;

II For $sign(\gamma_k) = -1$, with $\alpha > 1$ and $0 < p_k < \left(1 + \frac{2}{\alpha-1}\right)\pi_k$, when the polynomial (4.5) has exactly one positive root;

III For $sign(\gamma_k) = -1$ with $0 < \alpha \leq 1$, when the polynomial (4.6) has exactly one positive root.

So the algorithm for finding the optimal premium can be summarized in the following steps.

Step 1: *Collecting the necessary historical data from the insurance market.*

Firstly it is necessary to collect data useful for calculations as the number of companies which are in the market, the volume of business V_{k-1}, the break-even premium π_k, the impact of the reputation γ_k and the other stochastic parameters θ_k.

Step 2: *Estimation of parameter* α.

The parameter α is important because it indicates the elasticity of the market's average premium over the company's premium and it is estimated through the formula $\alpha = \frac{2\gamma}{\sigma^2} - 2\mu$, where γ is the coefficient of the non-linear damping, μ is the expected value of the white noise and σ is his standard deviation. These parameters are estimated with an appropriate data fitting.

Step 3: *Estimation of* $\bar{p}_k$.

The market's average premium is equal to $\bar{p}_{k-1} + \gamma \bar{p}_{k-1}^{2\mu-1} + \sigma \bar{p}_{k-1}^{\mu} \epsilon_k$, where $\{\epsilon_k\}$ is a sequence of uncorrelated normally distributed random variables with zero expectation and unit variance.

Step 4: *Calculation of* b_1 *and* b_2 *of the polynomial* (4.5) *and* (4.6).

Using the previous parameter α and the data collected in Step 1, coefficients b_1 and b_2 are estimated.

Step 5: *Calculation of the roots of the polynomial* (4.5) *and* (4.6).

The main possible directions for this step are related to the negative or positive effect of γ_k and this effect can be "measured" with the sentiment analysis, opinion polls etc. Consequently, when the impact of the reputation can be determined, the roots of the polynomials are calculated based on the previous Corollary.

Step 6: *Design and agreement on the optimal premium for the insurance company.*

In the last step, after taking into consideration the competition in the insurance market, the reputation of the company, the break-even premium, the different macroeconomic parameters (based on θ_k) and the elasticity parameters, the optimal premium can be calculated. This calculation must be repeated for different values of β: the β that represents the best scenario must then be chosen. Finally, this premium must be approved by the senior management of the companies.

If insurers want to maintain a competitive advantage in the insurance sector, however, they must not only take into account consumers and competitors, but they must adapt effectively to new technological complexities.

4.3 Pricing Strategy with the Arrival of New Technologies

In recent years, big data, the Internet of Thing and the predictive data analysis are heavily impacting on the pricing model: in fact, they allow to the insurance companies to estimate more accurately the risk or the consumer willingness to pay or buy, besides the fact that they help in identifying, during the underwriting phase, the insured who could commit fraud.

But the technology that has most altered the insurance landscape for both companies and consumers is Internet: in a relatively short time, it has become

the dominant channel of choice for people searching for insurance. There has been an increase in websites that allow consumers to compare insurance products by price, value and advantages and to match a product choice with their needs and willingness to pay, as if they were insurance brokers. All this has led the customers to be much more informed, sophisticated and open to new proposals, based on different variables (as security, mobility and different types of coverage), which require new and dynamic pricing structures. These new structures must consider several new factors.

Firstly, the increase in the use of Internet by consumers has led to a change in marketing by companies. In fact, it forces the companies to reappraise the traditional marketing channels, such as print and television, and to focus on *online marketing*.

Secondly, companies in the pricing strategy must take into account the fact that they can *no* longer rely on customer *loyalty*. The websites where there is a premium comparison attract the most "capricious" and price-sensitive customers, which are more inclined to revisit the price comparison site at renewal step. In contrast, the best way to capture a loyal customer was through personal recommendation o traditional media. For this reason, price comparison websites are themselves trying to create customer relationships and to build loyalty with the insurer. For example, many sites have strategies of cross-selling and communication that promote the benefits of going back to the website at renewal.

Thirdly, with the advent of the Internet the insurer brand and its *service benefits are more "denigrate"*: the companies are placed into a pot where premium is fundamental and the opportunity to compete based on non-price attributes is reduced. In other words, the choice of client becomes more dependent on price than benefits and added value services, typical of most traditional insurance, so the pricing strategy of company must be based primarily on the premium. Moreover, the increased premium sensitivity requires insurers to reduce profit margins: in fact, while previously the companies used to convince a customer giving him more benefits, now they are forced to be more competitive on the price and therefore to decrease the profit margin. In addition, the opportunity for insurer to increase premiums in the following years, with the aim to recover lost margins is reduced, because if the price is raised enough to increase profits,

the insurer risks to lose the customer.

Fourth thing, part of the company's strategy is based on understanding what are the *right questions* to ask the customer via the website. In fact, a wrong choice of questions can lead to incorrect premiums, that must then be corrected, leading to customer dissatisfaction, since the correction usually consists in increasing the premium.

The central questions are thus how to integrate websites with the overall marketing strategy of the company and how to obtain it without significantly eroding profitability. It is clear that, in order to have a good result, the company must take into consideration several factors, such as pricing anomalies, customer disloyalty, brand commoditisation and erosion of profitability. Further, the insurers who continue to rely only on a traditional actuarial model, with a perspective based on costs and a limited set of risk differentiators, will eventually end up with a larger pool of relatively riskier and less profitable customers: this will negatively impact on the profitability and the market share.

Conclusions

With this thesis writing a general overview on the theory and the main techniques related to the pricing in non-life sector, focusing on Car Liability insurance, is given.

Attention has been focused also on generalized linear models and in particular on the two models most frequently used by insurance companies in the pricing process, i.e. the models that assign Poisson distribution and gamma distribution to the response variables. The flexibility and peculiarity that characterize these models, make them widely used in the actuarial practice, not only for the construction of the fair premium *a priori*, but also for the evaluation of the premium reserves and, in some cases, for the determination of premiums *a posteriori*.

It can be concluded that for an insurance company it is fundamental to be able to estimate precisely what the expected costs will be for the customer and to understand, based on these, how much the maximum gain can be. Furthermore, the company must take into account various factors in order to be able to make the best profit. In primis, it must be considered that even if the Car Liability is a compulsory insurance, it cannot set too high prices due to the great competitiveness. Moreover, it must take into account the fact that the figure of the client has evolved: with the advent of the Internet, in fact, the customer has become more informed and sophisticated, making the competitiveness further increase among insurance companies.

To conclude, in order to continue to maximize profits, the company must continue to modify its pricing strategy, adapting it to the various new factors that will arise over the years.

Bibliography and sitography

ALTINI M. (2015), *Dealing with imbalanced data: undersampling, oversampling and proper cross-validation.*, https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation.

BCG (2019), *The Growing Importance of Pricing*, https://www.bcg.com/it-it/industries/insurance/growing-importance-pricing.aspx

BRANDIMARTE P., *Pricing Management: Motivation, History and Industrial Practice*, slides for Business Analytics - 2018/19, Dip. di Scienze Matematiche - Politecnico di Torino.

BRANDIMARTE P., *Pricing Management: Microeconomic Foundations*, slides for Business Analytics - 2018/19, Dip. di Scienze Matematiche - Politecnico di Torino.

DE JONG P., HELLER G. Z. (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press.

DRAKOS G. (2019), *Random Forest Regression model explained in depth*, Towards Data Science, https://towardsdatascience.com/random-forest-regression-model-explained-in-depth-f2cce437c750.

FAVARON F., *La tariffazione nelle assicurazioni contro i danni*, degree thesis discussed at the faculty of Economia e Finanza, Università Ca' Foscari Venezia, A.A. 2014/2015.

GIGANTE P., PICECH L., SIGALOTTI L. (2010), *La tariffazione nei rami danni con modelli lineari generalizzati*, Trieste, EUT Edizioni Università di Trieste.

HASTIE T. et al. (2013), *Introduction to Statistical Learning*, Ed. by Springer.

IVASS (2018), *Assicurazione R.C. Auto*, https://www.ivass.it/consumatori/normativa-consumatore/assicurazione-rc-auto/index.html.

MAYORGA W., TORRES D. (2017), *A practical model for pricing optimization in car insurance*, Panama, FASECOLDA.

OLIVIERI A., PITACCO E. (2011), *Introduction to insurance mathematics: technical and financial features of risk transfers*, Berlin, Heidelberg, Springer.

PASSALIDOU E., *Optimal premium pricing strategies for nonlife products in competitive insurance markets*, degree thesis discussed for the degree of Doctor in Philosophy, University of Liverpool, A.A. 2015.

PITACCO E. (2000), *Elementi di matematica delle assicurazioni*, Trieste, LINT.

PORZIO, et al. (2011), *Economia delle imprese assicurative*, Milano, McGraw-Hill.

ROBERTSHAW G. S. (2011), *Online price comparisons sites: How technology has destabilised and transformed the UK insurance market*, Quuensbury, BD13 1NR, UK.

SCHWARTZ A. J. (2015), *Price Optimization and Insurance Regulation With Examples and Calculations*, North Carolina Department of Insurance.

VERHOEF P. C., DONKERS B. (2001), *Predicting customer potential value: an application in the insurance industry*, Rotterdam, ERIM.

Acknowledgments

At the end of my university career, I would like to thank all the people who, in different ways, have been close to me during my studies and the writing of this Thesis.

I would like to thank first of all the Professors Vaccarino and Brandimarte, for the help and time they have dedicated to me during these months.

My gratitude also goes to RGI S.p.A. for the hospitality and the stimulating environment, that made my internship a personal and professional growth experience. A special thank goes to Enrica, Valentina and Valerio, that helped me with this Thesis.

A grateful thank also goes to my Family, that thanks to its economic and moral support has allowed me to reach this great goal.

Moreover, I would like to thank all my friends who have been present. In particular, Giulia, without whom my course of study would probably have been longer and more difficult, and Roberta and Sarah, who have been a source of valuable advice and inspiration for me. But also a heartfelt thanks to Marco, Paolo and Veronica.

Finally, I thank Simone who in the last two and a half years has been a good classmate and an even better life partner.