South East European Journal of Economics and Business
Volume 15 (2) 2020, 124-139
DOI: 10.2478/jeb-2020-0020

CLASSIFICATION RATEMAKING USING DECISION TREE IN THE INSURANCE MARKET OF BOSNIA AND HERZEGOVINA

Amela Omerašević, Jasmina Selimović

Abstract

This paper investigates the impact of risk classification on life insurance ratemaking with particular reference to Bosnia and Herzegovina (BiH). The research is based on a sample of over eighteen thousand insurance policies for passenger vehicles collected over the period 2015-2020. In our empirical investigation we develop a standard risk model based on the application of Poisson Generalized linear models (GLM) for claims frequency estimate and Gamma GLM for claim severity estimate. The analysis reveals that GLM does not provide a reliable parameter estimates for Multi-level factor (MLF) categorical predictors. Although GLM is widely used method to deter insurance premiums, improvements of GLM by using the data mining methods identified in this paper may solve practical challenges for the risk models. The popularity of applying data mining methods in the actuarial community has been growing in recent years due to its efficiency and precision. These models are recommended to be considered in BiH and South East European region in general.

Key words: *Risk premium, risk classification, generalized linear model, data mining methods, decision trees.*

JEL classification code: *G22*

1. INTRODUCTION

In a highly competitive economic environment, an important aspect for insurance business is the insurance premium, which the insured pays in exchange for the risk transfer to the insurer. One of the main tasks of actuaries in non-life insurance companies is ratemaking process for the portfolio of insureds, which will fairly distribute the risk among insureds. Risk classification is one of the most important elements in the ratemaking process of non-life insurance. Classification ratemaking is the process of grouping insurance policies of a certain insurance portfolio into homogeneous groups with similar expected claims experience or risk profile, so that all insureds within the same group pay the same insurance premium (Werner and Modlin 2010). Homogeneous groups of insureds created on the basis of risk classification are defined as risk classes. Insureds of a certain risk class generally have similar characteristics, which enable insurers to better assess the insurance premium. In general, the aim of risk classification is to determine a fair premium for each insured in the portfolio

Amela Omerašević, MSc
CFO
Uniqa osiguranje d.d. Sarajevo
E-mail: amela.omerasevic@uniqa.ba

Jasmina Selimović, PhD (corresponding author)
Associate Professor
School of Economics and Business
University of Sarajevo
E-mail: jasmina.selimovic@efsa.unsa.ba
Address: Trg Oslobođenja Alija Izetbegović 1
71000 Sarajevo, Bosna i Hercegovina

and to ensure the financial stability of the insurance company.

With the introduction of the III European Insurance Directive in July 1994, most insurance companies in the European Union (EU) countries took the opportunity to create their own insurance premium rates, based on statistical models for insurance ratemaking. Risk classification is very important in a competitive insurance market, where liberalization of prices is in effect. A better understanding of the real impact of risk classes on the insurance ratemaking can help insurance companies to improve their financial position, following the deregulation of the insurance market. Given that the deregulation of the motor third party liability insurance market in Bosnia and Herzegovina was initially planned for 2020 (but later postponed for October 2022 in the Federation of Bosnia and Herzegovina and for year 2023 for the Republika Srpska entity[1]), the development and application of statistical models for insurance premium ratemaking based on risk classification is a current topic for our insurance market as well. This paper aims to fill this gap by providing assessment between different methods of risk assessment and by reporting the method which should fit the BiH insurance market the best.

The paper is organized as follows. After introductory part, the next section provides relevant literature review of the field. The following section explains the methodology applied, including the Generalized linear models and the Decision trees data mining models, and explanation of the primary data used in the research. The obtained results are discussed in penultimate section, while the final section concludes the paper.

2. LITERATURE REVIEW

The most popular statistical models used today in the actuarial mathematics of non-life insurance are Generalized linear models (GLM). Credit for the development of GLM in both actuarial science and statistics belongs to British statisticians Nelder and Wedderburn (1972). They showed that GLM is an extension of traditional linear models, where the probability distribution of the dependent response variable is a member of the family of exponential distributions (normal, Poisson, gamma, ...), and the expectation of the dependent variable is determined using a linear predictor based on nonlinear link function. GLMs have become very popular and have proven effective in actuarial work over the last twenty years. The advantage of GLM over previously used methods of insurance ratemaking is the general statistical framework, which has established techniques for estimating standard error, confidence intervals, goodness of fit, etc. On top of that, standard statistical software (e.g. SAS, SPSS, R) for GLM makes the analysis of data to determine premium rates relatively easy.

Since the introduction of GLM to the present day, an abundance of outstanding work has been published, with many authors and scholars being able to highlight, develop or improve the assumptions that have enabled the practical application of these models in non-life insurance. Among the leaders of the GLM approach as the main statistical tool non-life insurance premiums ratemaking stands out McCullagh and Nelder (1989). Renshaw (1994) showed how GLM can be used to analyze the claims frequency and severity. Brockman and Wright (1992) used GLM to statistically model the frequency and severity for motor third party liability insurance. Haberman and Renshaw (1996) presented a comprehensive overview of the application of GLM to various actuarial problems in non-life insurance. Anderson et al. (2007) is the most useful guide for actuaries to apply GLM in practice and problem solving. Kaas et al. (2009) illustrated the usage of GLM for the bonus-malus system of motor third party liability insurance. Ohlsson and Johansson (2010) presented the basics of GLM theory for insurance ratemaking with an illustration of examples for multiplicative and hierarchical models, and useful extensions of GLM theory to Generalized Additive Models. Goldburd et al. (2016) published a comprehensive manual for GLM application in risk classification and tariff development in non-life insurance. Although actuaries are thought to have fully mastered GLM, improvements and enhancements of GLM for various applications in the insurance industry are still a hot topic (Hilbe 2014; Frees and Lee 2016; Garrido et al. 2016; Coskun 2016).

Insurance ratemaking process is a complex task, as it requires the development of a statistical model that should realistically show the impact of different predictors on insurance premiums. In the development of GLM it is necessary to include as many predictors as possible, which have an impact on the amount of insurance premium. However, the identification of the most significant variables and the correlations between them require a great deal of time for analysis. For certain categorical variables, with a large number of categories without a clear data order, there is no easy way to form groups with a sufficient amount of

[1] Note, in Bosnia and Herzegovina this field is regulated at the sub-national entity level. BiH is composed of two entities, the Federation of BiH entity and the Republika Srpska entity, and one district – the District Brcko of BiH.

data. The shortcomings of GLM are also the reasons for potential improvements of GLM for non-life insurance ratemaking application.

Actuaries have recently shown great interest in data mining, as data mining methods allow the analysis of large data sets that are common in many areas of insurance. In the last twenty years, data mining methods have become a useful tool in many areas of business, such as: marketing, financial services, investments, telecommunications, fraud detection, manufacturing and other areas of business. Some of the most well-known data mining methods are: Decision tree, Neural networks, Cluster Analysis, etc. Industries use data mining methods to achieve competitive advantage, increase efficiency, and provide better customer service (Fayyad et al. 1996). In recent years, the insurance industry has adopted the application of data mining methods as a strategic tool to compete in the insurance market. Data mining helps insurance companies in various business areas (SAS Institute 2000): insurance pricing, acquiring new clients, renewal of insurance portfolios, developing insurance products, detecting insurance fraud, reinsurance analysis, sophisticated marketing campaigns, claims assessment. A comprehensive overview of data mining methods presented by the authors Han, J. et al. (2012) and Hastie et al. (2001) systematically presented most of the statistical methods used in data mining today. Sumathi and Sivanandam (2006) explored the concepts of data mining and data warehousing, and presented areas of application in the insurance industry. Francis (2001) compared neural networks and regression models on insurance examples. Dugas et al. (2003) investigated the application of neural networks to determine motor insurance premiums in North America. Guo (2003) described the application of the decision tree method to model the claims frequency in non-life insurance. Shapiro and Jain (2003) presented a collection of papers by various authors on the theory and application of data mining methods in the insurance industry. Yao (2008) used cluster analysis methods to determine the claims frequency by geographical areas. Derrig and Francis (2006) presented the application of the C&RT decision tree to address insurance fraud detection. The work of Kolyshkina et al. (2004) discusses the advantage of combining GLM with a multivariate adaptive regression mining method. Williams et al. (2015) compared different data mining methods for the selection of predictors on the example of a property insurance premium.

Data mining uses a variety of data analysis methods to research data and discover useful patterns and trends. Models developed using data mining methods are more accurate, faster, and more efficient at solving business problems. Data mining methods can be combined with GLM to improve GLM prediction performance and/or efficiency. Data mining methods are useful in overcoming the shortcomings of GLM, because they help in analyzing large amounts of data, searching for hidden patterns in the data and obtaining useful information. One possibility of applying data mining methods combined with GLM in insurance premium ratemaking is to reduce the number of categories in categorical variables, by eliminating categories with insufficient data, to leave only significant categories for predicting the response variable. This paper argues that the data mining methods and GLM can be combined to take advantage of both approaches.

3. *METHODOLOGY AND DATA*

The risk premium approach traditionally has been used to determine non-life insurance premiums, so it will be applied in this research as well. Risk premium is the average expected amount of claims under the insurance policy during the insurance period, i.e. the period of risk exposure. Risk premium represents the expected amount of all claims reported by the insured during the insurance period, and is obtained by multiplying the two components, the expected value of the claims frequency and the expected value of the claims severity. The claims frequency is the number of claims incurred per policy during the period of risk exposure. The claim severity is the total claim amount divided by the number of claims incurred during the insurance period. Due to the previously mentioned advantages of a separate assessment of the claims frequency and severity, two separate statistical models will be developed:

- GLM for claims frequency estimate and
- GLM for claims severity estimate.

These two models will be combined to obtain a model for risk premium, called the standard risk model. GLM estimates expected claims frequency and expected claims severity based on key predictors or risk factors. Risk factors are the characteristics of the insured, the subject of insurance and his environment that are believed to directly affect the claims frequency and the claims severity during a given insurance period.

The standard approach in the insurance industry for the selection of risk factors in GLM is based on the statistical significance of the predictors for the response variable. Risk classes for categorical predictors in GLM are determined by grouping categories with insufficient risk exposure together. Categories with sufficient risk exposure are considered separately.

In practice, very often there categorical variables used as risk factors have a large number of categories or „levels", without sufficient amount of data in each category. Ohlsson and Johansson (2010) introduced the notion of Multi Level factor – MLF for such categorical variables. In this case, it is necessary to group the categories before joining to GLM. Otherwise, there will be no convergence of the model. When MLF variables are included in the GLM, a large number of parameters with a significant standard error are obtained. On the other hand, if these variables are not included in the GLM as predictors, we encounter the problem of over-parameterization of the model. Therefore, for MLF categorical predictors, it is necessary to reduce the number of categories/levels of such variables, i.e. to apply to them some of the methods of reduction cardinality for categorical variables. Reduction of cardinality for categorical variables, whether nominal or ordinal, is the process of combining two or more categories into one new category.

In order to improve the accuracy of risk premium prediction, this paper investigates data mining methods that will compensate for the shortcomings of the standard risk model, in terms of determining risk classes for MLF predictors. In the following text, for the simplicity sake, the term predictive model will be used for the risk model that combines GLM with data mining methods. The application of data mining methods in the predictive model aims to reduce the cardinality of MLF categorical variables, i.e. optimal grouping of categories with minimal loss of information. Given that there are a large number of mining methods for the selection of risk classes, this method investigates the decision tree methods. As a result of the selection of risk classes using data mining methods, new predictors will be obtained, which will be used as input parameters in GLM for claims frequency and claims severity estimates of predictive model.

3.1 Generalized linear model

The first illustration of the application of GLM to insurance premiums pricing was presented by McCullagh and Nelder in their work (1989) on the example of estimating the average amount of claim in motor insurance. The purpose of GLM is to estimate the dependent response variable, which we denote by Y based on a number of known independent variables $\mathrm{X_i}$, where $\mathrm{i} = 1, .., \mathrm{n}$. To determine the non-life insurance premium, the response variable Y can be one of the following variables: number of claims, claims frequency, claims severity, risk premiums. The independent variables $\mathrm{X_i}$ are called predictors. Potential forecasting variables are the characteristics of: insurance policies, insureds and insured subject, which have an impact on the response variable. Predictors in GLM can be: categorical variables and continuous variables. The following three components need to be defined for the GLM specification:

(GLM1) Random component: The response variable Y belongs to the exponential family of distributions (normal, Poisson, gamma, binomial, exponential, etc.), if its density can be written in the form:

$$f_Y(y_i;\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\},$$
$$i = 0, \dots, n;$$

Poisson distribution and negative binomial distribution are most commonly used to model the claims frequency or the number of claims. Gamma and inverse Gaussian distributions are best suited for modeling the amount of claims or claims severity, due to the positive values of the response variable. The parameter θ_i is related to the mean $\mu_i=E[Y_i]$. The scale parameter ϕ is taken as a positive fixed value or estimated from data based on Pearson's moment estimator or using the maximum likelihood method. The different choice of functions $a(\bullet)$, $b(\bullet)$ and $c(\bullet)$ specifies a distribution function, suitable for solving GLM problems. The dispersion function $a_i\ (\phi)$ often has a form $a_i\ (\phi)=\phi/\omega_i$, where ω_i are prior weights of the exposure of the i-th observation. In modeling the claims frequency, the prior weights are equal to the risk exposure, and in modeling the amount of claims, prior weights are equal to the total number of claims.

(GLM2) Systematic component: Linear predictor η_i is a linear function of independent predictors X_{ij} and unknown parameters β_j:

$$\eta_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \xi_i, \quad i = 1, \dots, n$$

where: n is the number of known data that are the subject of observation, p is a number of model parameters, $n-p$ degree of freedom, X_{ij} predictors, β_j model parameters and ξ_i offset. The values of the parameters β_j are estimated by the method of maximum likelihood. For each level of the categorical prediction variable, in addition to the base level in GLM, one parameter is determined in the linear predictor. Only one parameter in a linear predictor is determined for each continuous prediction variable. If all categories of one categorical prediction variable are included in a linear predictor, then such a variable is called the main effect. An interaction between two categorical predictors can be included in the model, which allows the influence of one prediction variable on the response variable to depend on the value of the other prediction variable. Interactions are included in the model only if they are statistically significant and if their inclusion is justified to estimate the response variable. One of the main advantages of GLM is the ability to determine the reliability of the estimation of the parameter β_j. To determine the significance of a predictors standard GLM software tools provide several parameter diagnostics, which can help in the reliability of parameter estimation, including: standard error, confidence interval, statistical tests.

(GLM3) Link function: The relationship between a random and a systematic component is defined through a link function $g(.)$, which is a differentiable and monotonic function via the equation:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Estimates of the mean μ_i are obtained by applying the inverse function $g(.)$ to the obtained values of the linear predictor:

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

Link function $g(.)$ provides GLM flexibility in defining the relationship between mean and the linear predictor. The flexibility to use different link functions gives more opportunities to specify a model that better reflects reality. In theory, different link functions can be used for different estimates, but this is rarely applied in practice. The choice of the link function depends on the choice of the response variable that we estimate in the model. The Log link function is most often used to determine the insurance premium, i.e. $g(x)=ln(x)$. GLM with Log link function has the property to produce multiplicative models. By using the Log link function, the sum of the components of the linear predictor is converted into the product of the components of the linear predictor, i.e. the additive model is converted into a multiplicative model. The multiplicative model is a widely used model in insurance ratemaking, due to the advantages that its structure offers for the development of premium tariffs. Multiplicative models are simpler and more practical to apply than additive models, and the premium is always a positive value, without any additional adjustments.

3.2 Decision trees data mining methods

Decision trees are a very popular data mining method, used to solve classification and prediction problems. The advantages of the decision tree over statistical methods or other data mining methods are as follows:

- Decision trees are built quickly and easily, and give simple and understandable results.
- Graphical presentation of data in the form of a tree helps to understand the causes and impact of predictors on the response variable.
- Decision trees can also work with variables that have missing values in the data.
- The collinearity of the predictors does not affect the performance of the decision tree.
- Most decision tree algorithms are very fast and for a large amount of data.
- Decision tree models are particularly suitable for problem solving when there is no a priori information about the relationship function between the predictor and the response variable.
- Decision trees give very accurate prediction results.

Decision trees do not provide a large number of diagnostic tests and statistical measures, which allow statistical methods such as GLM. Furthermore, if there is a linear relationship between the predictor and the response variable, then the statistical methods have better performance than the decision tree. On the other hand, decision trees do not have special constraints or requirements if used in the data preparation process, for later use in a statistical model such as GLM.

If a decision tree is being built for a large amount of data, then it is necessary to apply an algorithm to build a decision tree. Most decision tree construction algorithms use a recursive top-down partitioning approach. A large number of researchers from various scientific and technical disciplines have dealt with the problem of induction of the decision tree based on available data. The first decision tree algorithm known as ID3 (Iterative Dichotomiser 3) was developed by Australian researcher Quinlan (1986). Quinlan (1993) introduced the new algorithm C4.5, which has become a benchmark against which newer decision tree algorithms are compared. Breiman, Friedman, Olshen, and Stone (1984) described the generation of binary Classification and Regression Trees (C&RT). The decision tree algorithms ID3 and C&RT were developed independently at about the same time, and yet follow a similar approach to induction, i.e. decision tree growth. These two fundamental algorithms have prompted a series of studies based on decision tree induction. Several decision tree algorithms are used today: ID3, C4.5, C&RT, CHAID, QUEST, and RF. Each of these algorithms has unique qualities in building a decision tree. For data classification, in which the dependent variable is categorical, all algorithms can be used to build a decision tree. For regression problems where the dependent variable is continuous, only CHAID and C&RT can be used. Given that the claims frequency and the claims severity are continuous response variables, the decision trees of CHAID and CR&T for determining risk classes are considered in this paper.

The CHAID decision tree algorithm is one of the oldest decision tree methods originally proposed by Kass (1980). The CHAID decision tree method is also one of the oldest data extraction methods and one of the earliest to appear in the actuarial literature (Gallagher et al. 1990). CHAID is an acronym for x^2 automatic interactive detector. CHAID decision trees recursively divide data into two or more groups so that the data in each group is more homogeneous than previously divided data. CHAID uses categorical and continuous predictors, with continuous predictors divided into a number of categories, with approximately the same number of records, before a decision tree is built. The next step is to go through all the predictors, for each prediction variable determine a pair of categories that differ the least from the response variable, using the x^2 test for classification problems and the F test for regression problems. If the corresponding test for a particular pair of variable predictors is not statistically significant relative to the defined value of a, then the respective categories of the predictor variable will be merged into a node and this procedure will be repeated for the next pair of categories. If the minimum p-value for each prediction variable is greater than some default value of a, there are no further divisions of the decision tree and that node becomes a terminal node. The process continues until there is a further division of the decision tree.

C&RT is an abbreviation for classification and regression decision trees, originally described in Breiman et al. (1984). This paper describes the general theory of classification and regression decision trees and presents specific solutions for building a C&RT decision tree algorithm. C&RT uses a binary recursive approach to building a decision tree, so that the data set is divided into exactly two subsets, with the records within each subset being more homogeneous than in the previous subset. Each of these two subsets is again divided, and the process is repeated until the criterion of homogeneity is reached or until some other criterion of stopping the decision tree is met. To solve classification problems, a C&RT tree can be constructed using various node division criteria and the most commonly used measures are Gini index, x^2 or G^2. To solve regression problems, the C&RT decision tree for node division uses the method of least squared deviations.

3.3 Data

Changes in legislation related to the liberalization of the motor third party liability insurance market, which should first enter into force by the end of 2020 (but this deadline was extended for end of 2022 and 2023 in the Federation of BiH and the Republika Srpska entity, respectively) will inevitably affect the reduction of the insurance market in Bosnia and Herzegovina. According to the experiences of neighbouring countries Croatia and Slovenia, two years after the introduction of liberalization of prices the motor third party liability insurance market have decreased by 30%, which has resulted in significant losses to the insurance industry in these countries. Insurance companies partially compensated for losses in motor third party liability insurance by adjusting the insurance premium to motor hull insurance, as well as by introducing new additional insurance products. Although the motor hull insurance market in BiH is much smaller compared to the motor third party liability insurance market, it is at the same time much more flexible and exposed to changes inside and outside the insurance industry. The sensitivity of clients to price changes in motor hull insurance is higher compared to all other insurance products, while customer loyalty is very low. If an insurance company does not have an adequate risk classification, it is likely to be subject to anti-selection. This means they will offer low prices for high risks and high prices for low risks. Better risks will leave the insurance company, attracted by the lower premium of competitors, which will lead to further financial loss of insurance companies. The best way to avoid anti-selection is to make a more accurate estimate of the insurance premium, based on the risk classification. Due to the aforementioned reasons, the research was conducted on the development of models for premium rate making of motor hull insurance in Bosnia and Herzegovina. Generalized linear models were used for risk premium ratemaking and risk classification was done using data mining methods. The insurance data of motor hull insurance of one of the leading insurance companies in Bosnia and Herzegovina were used for the research. Insurance data from the last 5 consecutive years are a good basis for model development. The sample consists of 18,012 insurance policies for passenger vehicles of the insured and data on the history of claims. In the case of motor hull insurance, compensation is paid to the insured in case of damage, or loss of vehicles and / or equipment as a result of the following insured hazards: traffic accidents, burglary, fire, lightning, explosion, fall and impact, storm, hail, snow, avalanches, floods and torrents, aircraft crashes, demonstrations, malicious actions by third parties and broken glass.

Based on the available information, 25 potential input predictor variables were selected for model development, which are grouped into three categories: characteristics of the insured object, i.e. vehicle, characteristics of the insurance policy and characteristics of the insured. To create an initial set of insurance policy data, vehicle records are associated with insurance policy records. One insurance policy record corresponds to the time period during which the vehicle was exposed to the insured event and the risk of damage. To create the initial set of claims data, the claim records are presented in such a way that one claim request refers to one insurance event to which the vehicle is exposed. To integrate data from previously created data sets on insurance policies with the set of data on claims, for all claims under the insurance policy, only one record was created in the database by summing the number of claims and the total amount of those claims.

Certain data mining techniques can only work with variables that have numerical data values. For this reason, the transformation in the values of nominal variables into numerical values was performed. After data preparation, the final data set for modeling the claims frequency and claims severity was formed, which contains 22 variables and 17,404 records on motor hull insurance policies. Due to the relatively small data set used in this study, the data were divided by random distribution: 80% training data set for model development, and 20% test data set for model testing and evaluation.

4. RESULTS AND DISCUSSION

This section presents a standard approach of risk model development with GLM for claims frequency and GLM for claims severity estimate in the manner common in actuarial practice. The standard risk model was used as a reference for comparison with predictive models, which include CHAID and C&RT decision tree data mining methods for selecting risk classes for MLF categorical predictors. Models for CHAID and C&RT decision trees have been developed and as a result a new predictor with a smaller number of categories or risk classes has been obtained. The new predictors, obtained on the basis of the mentioned data mining methods, were used as input variables in GLM for the claims frequency and GLM for the claims severity. Based on the model assessment, the best decision tree method for risk classification was selected. Finally, an evaluation of the standard risk model and the best predictive risk model was performed taking into account business objectives and key performance indicators.

4.1 Standard risk model

Standard GLM models for claims frequency and claims severity were created in the following five steps:

1) The parameters of the model are defined: distribution function, link function, response variables and predictors;
2) The significance of each prediction variable was tested, as well as the significance of interactions between predictors;
3) A model is formed on the basis of significant forecasting variables;
4) The significance of each of the parameter estimates was tested;
5) The final GLM model was formed.

GLM for claims frequency estimate

The Poisson distribution was used to model the claims frequency as the most popular distribution for modeling the claims frequency in non-life insurance. Namely, the first choice for modeling the claims frequency or the number of claims in the literature is GLM with Poisson distribution, according to Antonio and Valdez (2010), Dionne and Vanasse (1988, 1992), Denuit and Lang (2004), Flynn and Francis (2009). Although GLM with Poisson distribution can also be applied to continuous response variables, due to the simplicity of the model, the number of claims for the response variable was used as a continuous variable instead of the claims frequency. Since not all insurance policies in the data set have the same risk exposure, the Log (*Exposures*) is included in the model as an offset when calculating the number of claims. For the link function, the canonical link function, $g(x)=ln(x)$, was used to make the model multiplicative. The scale parameter ϕ for the Poisson distribution is equal to 1.

All predictors are included in the GLM, to obtain the model with the best Akaike information criterion - AIC. The predictors are shown with the original records of the training data set at the insurance policy level. For nominal variables, the category with the highest risk exposure was taken as the base risk class, while for ordinal variables the smallest category was taken as the base risk class. For the hypothesis test of predictors significance, the Wald test and Type III analysis were used to determine the variables to be retained in the model. The Wald test follows a x^2 distribution with a statistically significant value of $p \leq 0.001$ and df degrees of freedom. After excluding from the model all predictors where the p-value is greater than the statistically significant value, the remaining variables (all variables are explained in Appendix 1, Table A1): *MarkaD, KlasaD, NamjenaD, LeasingD*, *Trajanje_ugD* and *Tip_osigD* are statistically relevant, which clearly indicates their impact on claim frequency. The statistical significance of the parameters of each predictor was checked and the results are satisfactory for the risk factors *NamjenaD, LeasingD, Trajanje_ugD* and *Tip_osigD*, because these are categorical variables with low cardinality, where enough records is included in each category. Figure 1 shows the estimated value of the *LeasingD* prediction variable with a 95% confidence interval. *LeasingD* satisfies the so-called horizontal line test, i.e. the horizontal line cannot be drawn in the confidence interval between categories 0 and 1. Both categories of the *LeasingD* predictor are statistically significant.

However, for MLF categorical predictors *MarkaD* and *KlasaD*, GLM results show a high degree of uncertainty. *MarkaD* is a significant variable for claims frequency, but 29 categories of this variable show a high standard error. The reason for this is the fact that a many level of *MarkaD* do not have enough risk exposure, i.e. enough insurance policies to be included in GLM as parameters. Figure 2 shows the 95% confidence interval for the *MarkaD* risk factor.

The confidence interval is wide for low-exposure categories, so these parameters of the *MarkaD*

Figure 1. 95% confidence interval for LeasingD variable

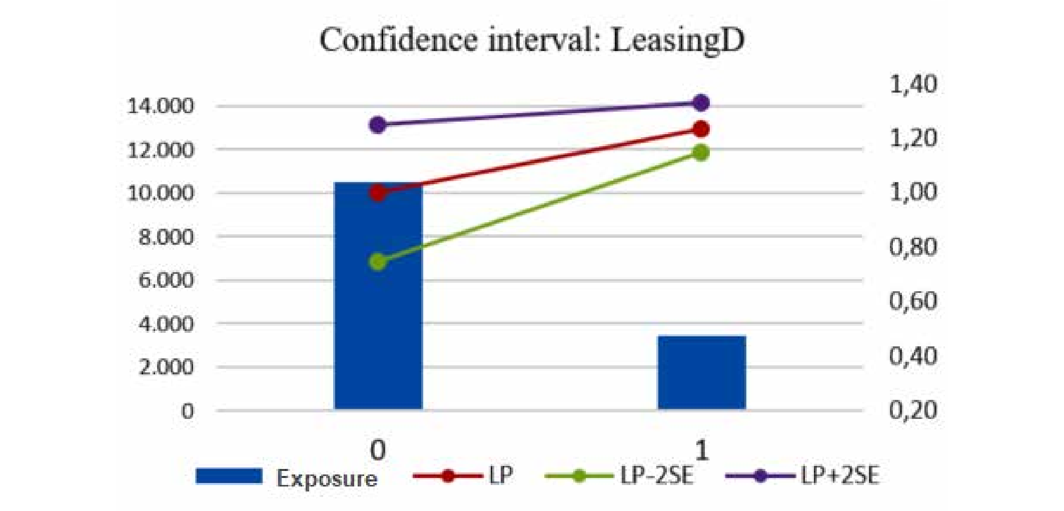


Figure 2. 95% confidence interval for MarkaD variable

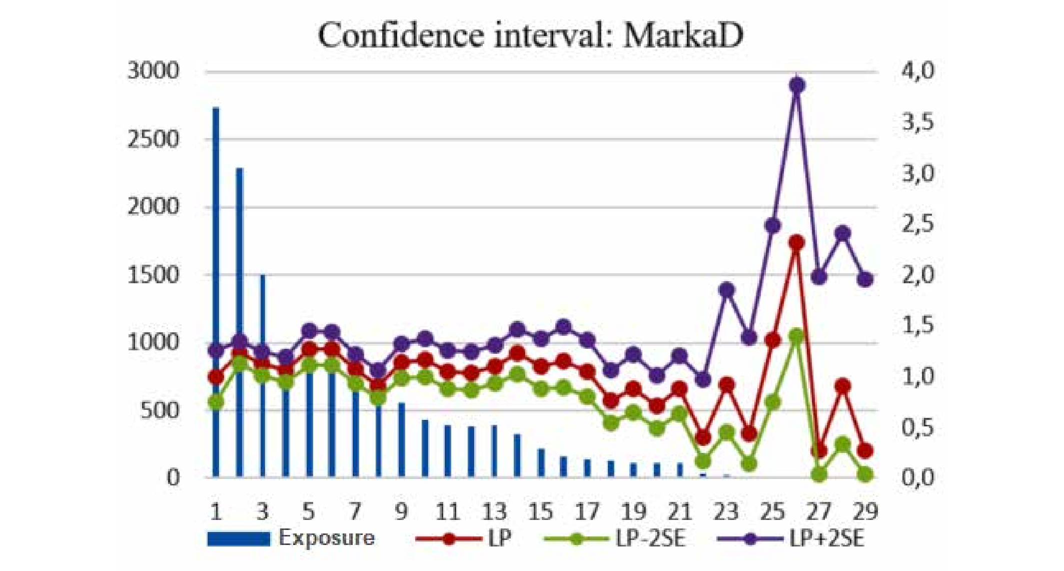


Table 1. Poisson GLM: standard approach

Parameter	B	SE	Hypothesis Test		
			Wald χ2	df	Sig.
Intercept	-0.82	0.02	1261.31	1	0.00
NamjenaD=1	-0.16	0.03	22.56	1	0.00
NamjenaD=0	0.00				
LeasingD=1	0.25	0.04	45.69	1	0.00
LeasingD=0	0.00				
Trajanje_ugD=1	-0.26	0.05	30.76	1	0.00
Trajanje_ugD=0	0.00				
Tip_osigD=1	0.22	0.04	33.21	1	0.00
Tip_osigD=0	0.00				
MarkaT=4	0.74	0.26	8.15	1	0.00
MarkaT=3	-0.98	0.45	4.78	1	0.03
MarkaT=2	0.16	0.06	8.85	1	0.00
MarkaT=1	0.00				
KlasaT=4	0.28	0.06	19.99	1	0.00
KlasaT=3	0.07	0.03	5.36	1	0.02
KlasaT=2	-0.15	0.03	18.72	1	0.00
KlasaT=1	0.00				

predictor will not pass the horizontal line test and therefore they are not statistically significant for the claim's frequency.

An alternative solution to this problem is to group categories that do not have enough exposure. On this way, "new" categories are created, i.e. risk classes with higher exposure, which can give credible results with GLM. The standard approach to grouping categorical variables is to add risk classes that are not statistically significant to the underlying risk class, in order to obtain customized predictors with sufficient exposure. In this case, the original *MarkaD* and *KlasaD* predictors have been replaced with the new *MarkaT* and *KlasaT* categorical variables, which have a smaller number of categories. All parameters for selected risk factors and created risk classes are statistically significant (p<0.05), which is satisfactory for calculating the expected values of the claims frequency.

The scale parameter $\hat{\phi} = 0.92$, obtained based on the total deviation is less than 1, which shows that the variance is less than expected, and it can be concluded that Poisson distribution is adequate for claims frequency estimate.

GLM for the claims severity estimate

As in the actuarial literature (Ohlsson and Johansson 2010; Parodi 2014; Kaas et al. 2009) gamma distribution is the most common distribution for modeling the claims severity and the natural choice for GLM claims severity estimate. The Gamma distribution variance function assigns greater variance to higher-expected claims, which is a desirable property when modeling the claims severity in GLM, even when the scale parameter ϕ is constant for all claims. To achieve the multiplicative model, instead of the canonical link function, the log link function, $g(x)=ln(x)$, was used. Pearson's moment estimator was taken as the scale parameter ϕ, as this approach was used by McCullagh and Nelder (1989) to obtain a more conservative estimate of variance. Most predictors are not statistically significant and have no effect on the average amount of claim. All variables with a p-value greater than 0,001 according to the Wald test, were excluded from the model. The results obtained based on the type III analysis for gamma GLM (Table 2) suggest that the claims severity to the analyzed portfolio is only affected by

Table 2. Gamma GLM: standard approach

Parameter	B	SE	Hypothesis Test		
			Wald χ2	df	Sig.
Intercept	7.56	0.05	23299.62	1	0.00
Lojalnost=2	-0.28	0.07	18.21	1	0.00
Lojalnost=1	-0.08	0.06	1.64	1	0.20
Lojalnost=0	0.00				

predictor Lojalnost. Risk factor that affect the claims severity differ significantly from risk factors that affect the claims frequency, which confirms the assumption of the actuarial literature on a separate analysis of these two phenomena.

4.2 Predictive model

This chapter discusses the CHAID and C&RT decision tree methods for the risk classification of MLF categorical variables. CHAID and C&RT models have been developed for MLF predictors: *MarkaD, KlasaD, OpcinaD* and *Osig_sumaD*. New predictor variables have been formed with a smaller number of categories, i.e. risk classes, which have sufficient risk exposure in each class. The new variables were obtained based on the decision tree method and were included in the Poisson GLM for the claims frequency and the Gamma GLM for the claim severity.

A comparison of GLM with risk classes selected with a standard approach and data mining methods was performed. Criteria for ranking and selection of the best GLM are:

- Goodness of fit,
- Prediction performance of the model.

Data mining models for risk class selection and GLM were developed on a training data set. Comparisons of GLM performance were performed on a test data set. The Akaike Information Criterion (AIC) was used as a measure of the goodness of fit. Information criteria represent the ratio between the accuracy of model adaptation to data and the complexity of the model. The lower the information criterion, the better the model is considered. The Gini coefficient on the test data set was used for the reliability of the parameters estimate. The Gini coefficient, named after statistician and sociologist Corrado Gini, is commonly used in economics to measure national income inequality. Meyers (2007) introduced the Gini coefficient as a general procedure to assess model prediction performance. The Gini coefficient does not quantify the profitability of a particular risk model but determines the model's ability to segment the best and worst risks. The higher the Gini coefficient, the better the model prediction performance is considered.

Selection of risk classes

To illustrate the results of data mining methods for classification the diagrams of the CHAID and C&RT decision tree for the *KlasaD* predictor are shown in Figure 3 and Figure 4, respectively. Given that the MLF variable *KlasaD* has an impact only on the claims frequency, in decision trees the claims frequency was used as the response variable. Terminal nodes contain the final prediction of the model and the belonging of each

Figure 3. *CHAID for prediction variable KlasaD*

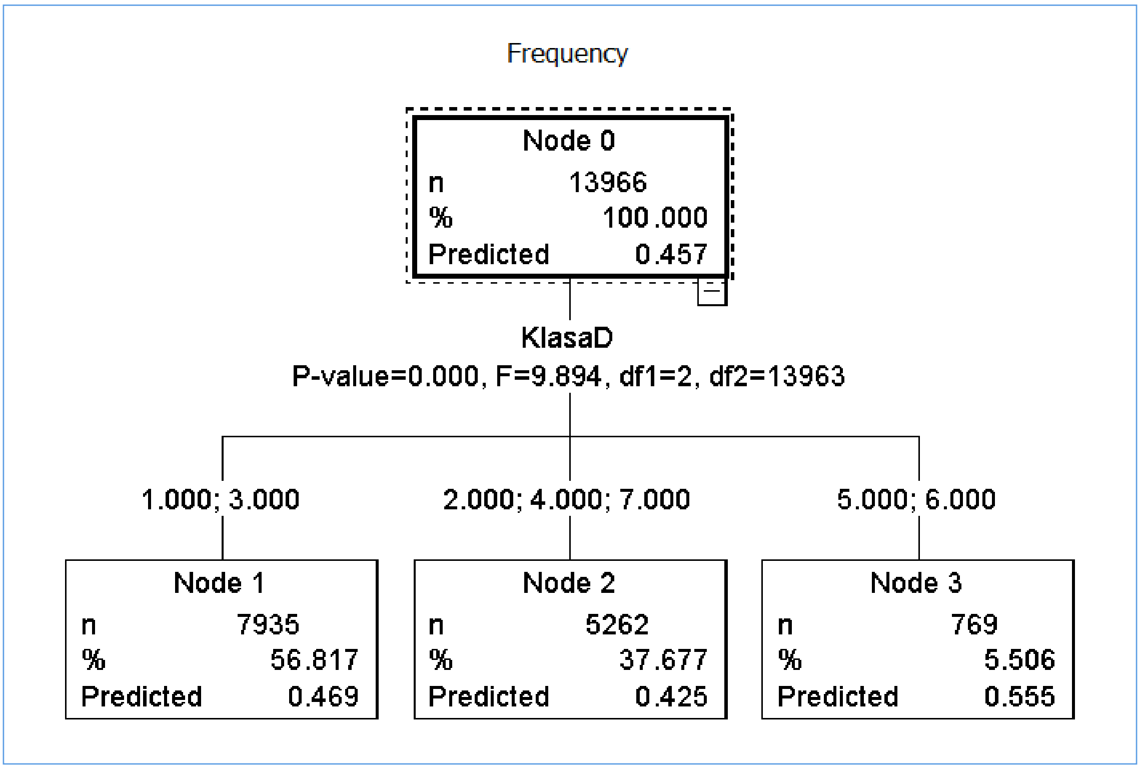
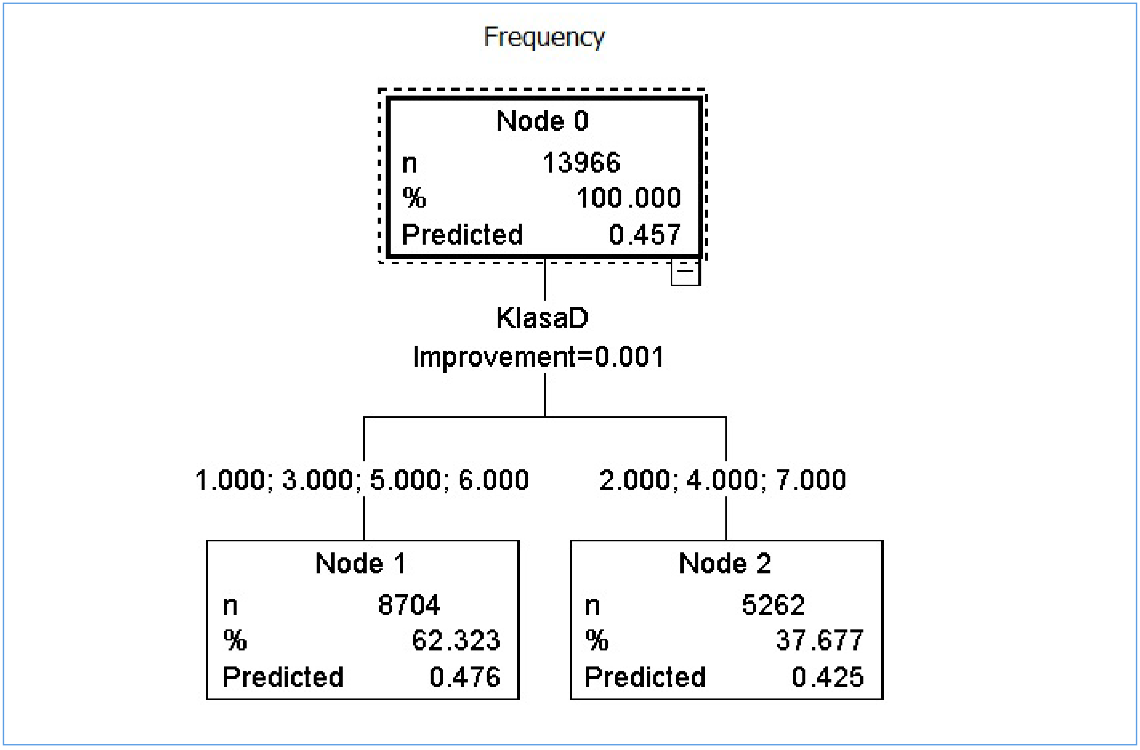


Figure 4. C&RT for prediction variable KlasaD



category to one of the risk classes. The new category variable *Klasa_CHAID* with three risk classes, created using the prediction results of the CHAID decision.

By applying the CR&T decision tree, the number of categories for the *KlasaD* prediction variable was reduced from seven to two risk classes, and a new predictor *Klasa_CRT* was obtained.

GLM for claims frequency estimate

The AIC measures to compare the Poisson GLM claim frequency model are shown in Figure 5. The lowest AIC has the GLM for claims frequency estimate with risk classes determined for MLF variables based on the CHAID decision tree, followed by the model with risk classes obtained using the CR&T method.

Gini coefficients of Poisson GLM for claim frequency estimate shown the CHAID decision tree achieves the best results in selecting risk classes for the claims frequency.

The parameters estimate and their standard errors for Poisson GLM claims frequency with risk classes selected for MLF predictors with CHAID decision tree are shown in Table 3. The selected predictors and their parameter estimates are statistically significant and have an impact on the claims frequency. The selection

Figure 5. AIC – Claim frequency model ranking

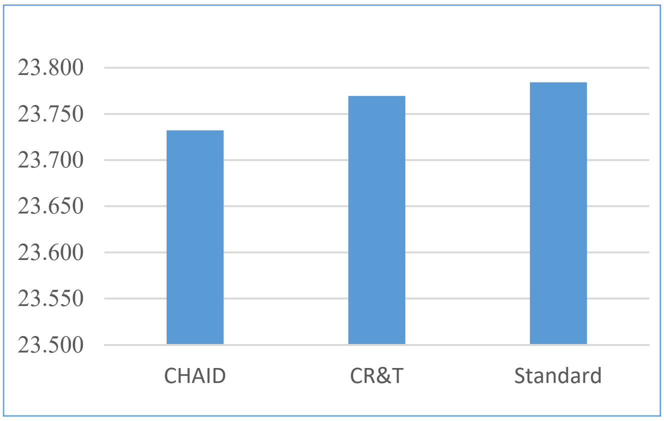


Figure 6. Gini - Claim frequency model ranking

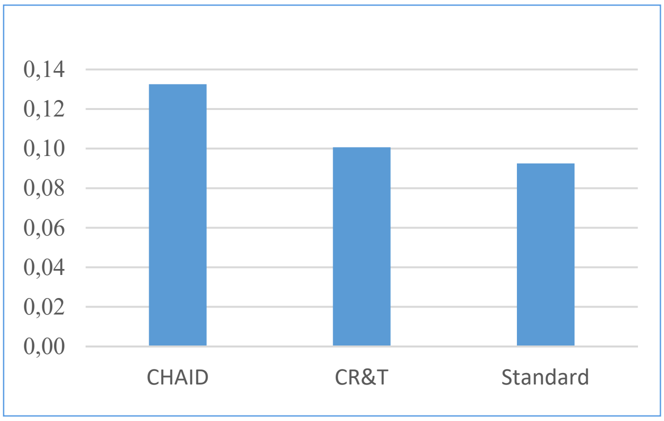


Table 3. Poisson GLM: CHAID risk selection for MLF

Parameter	B	SE	Hypothesis Test		
			Wald $\chi 2$	df	Sig.
Intercept	-1.24	0.05	552.26	1	0.00
NamjenaD=1	-0.16	0.03	23.57	1	0.00
NamjenaD=0	0.00				
LeasingD=1	0.21	0.04	27.01	1	0.00
LeasingD=0	0.00				
Trajanje_ugD=1	-0.26	0.05	29.01	1	0.00
Trajanje_ugD=0	0.00				
Tip_osigD=1	0.20	0.04	24.20	1	0.00
Tip_osigD=0	0.00				
Marka_CHAID=4	0.31	0.05	45.77	1	0.00
Marka_CHAID=3	0.20	0.05	16.66	1	0.00
Marka_CHAID=2	0.14	0.05	9.25	1	0.00
Marka_CHAID=1	0.00				
Klasa_CHAID=4	0.32	0.06	27.56	1	0.00
Klasa_CHAID=3	0.18	0.04	27.18	1	0.00
Klasa_CHAID=2	0.15	0.04	16.81	1	0.00
Klasa_CHAID=1	0.00				
Opcina_CHAID=3	0.23	0.05	18.82	1	0.00
Opcina_CHAID=2	0.13	0.03	17.88	1	0.00
Opcina_CHAID=1	0.00				

of optimal risk classes using the CHAID decision tree method before inclusion in GLM, improves the predictive performance of the claim frequency model.

GLM for claims severity estimate

AIC measures shown that application of any decision tree method for the risk selection achieves better results for the claims severity compared to the standard approach.

Comparing the Gini coefficients for claim severity estimate shows that GLM gamma with risk classes based on the CHAID decision tree gives the best results, compared to all other models used.

The selection of the optimal number of risk classes using the CHAID method has effect on the GLM for the claims severity. The parameters estimate from Table 4 shows that each risk class created based on the grouping of MLF predictors using the CHAID method is also statistically significant.

Figure 7. AIC – Claim severity model ranking

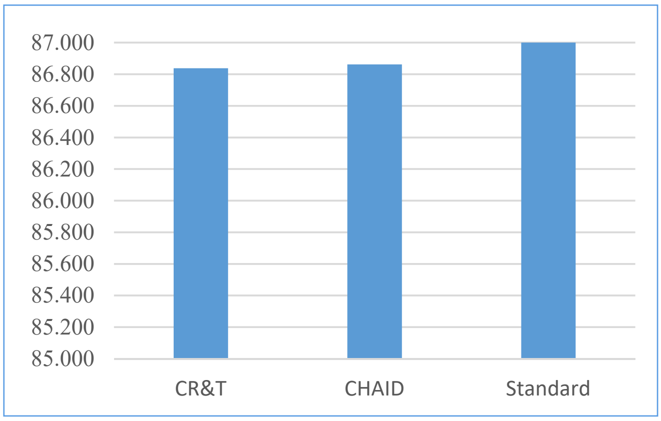


Figure 8. Gini - Claim severity model ranking

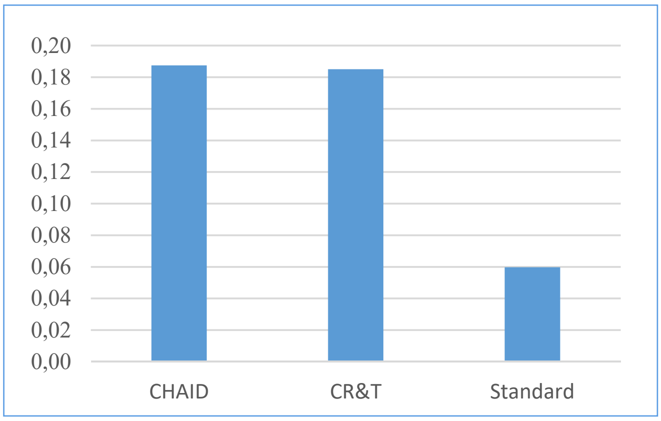


Table 4. *Gamma* GLM: CHAID risk selection for MLF

Parameter	B	SE	Hypothesis Test		
			Wald χ2	df	Sig.
Intercept	7.47	0.06	18102.16	1	0.00
NamjenaD=1	0.16	0.04	14.66	1	0.00
NamjenaD=0	0.00				
Lojalnost=2	-0.26	0.05	22.49	1	0.00
Lojalnost=1	-0.04	0.05	0.48	1	0.49
Lojalnost=0	0.00				
OS_CHAID=5	0.55	0.09	35.80	1	0.00
OS_CHAID=4	0.21	0.08	6.31	1	0.01
OS_CHAID=3	-0.43	0.05	72.03	1	0.00
OS_CHAID=2	-0.18	0.06	9.29	1	0.00
OS_CHAID=1	0.00				

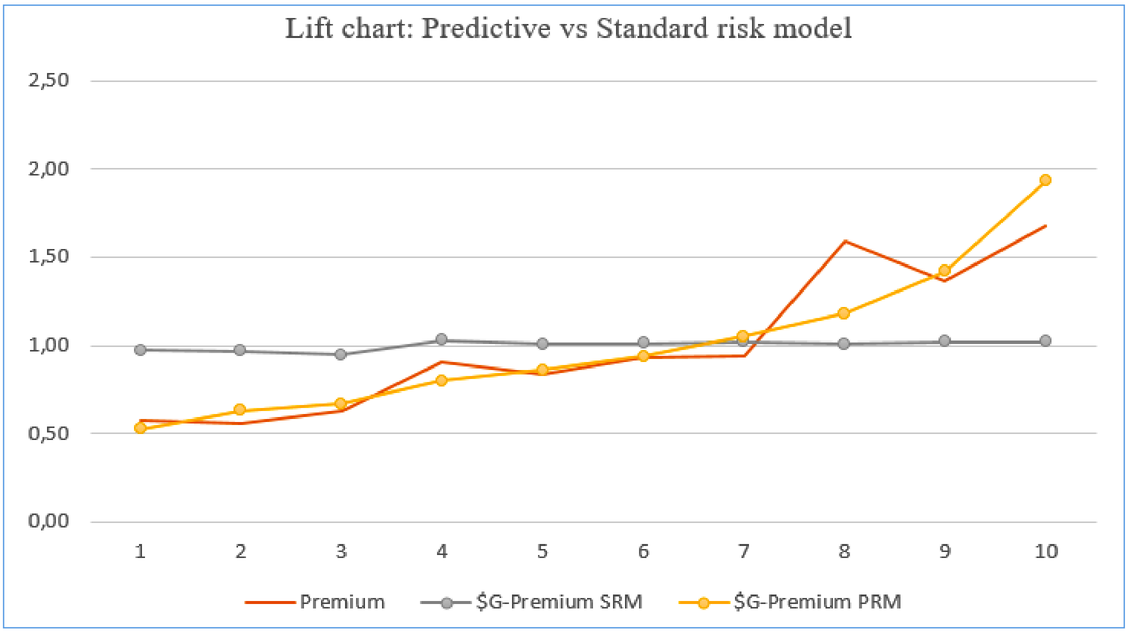
From the assessment of the claim frequency and claims severity models, it can be concluded that the CHAID decision tree shows excellent performance and should be applied to the selection of risk classes. Given the ease of application and speed of development of the CHAID model, the time required to select risk classes can be reduced by using this method. Predictive models that combine GLM for with the CHAID decision tree for risk class selection are more accurate and achieve better assessment results.

4.3 Model evaluation

An evaluation of the standard risk model with the best predictive risk model was performed in this section. The standard risk model was created based on standard models for claims frequency and claims severity. For the best predictive risk model, a model was chosen that that combine GLM for with the CHAID decision tree for risk class selection. The risk of premium rates, due to the application of the multiplicative model, is obtained by multiplying the relativity of the claims frequency and the relativity of the claims severity.

For comparison of the economic value of the two risk models the Lift Chart is used (Tevet 2013). The Lift chart helps to visually quantify the model's capability by charging fair prices to insureds. To visually compare the predictive and standard risk models, a double Lift graph was used, in Figure 9, on the test data grouped into deciles. The records are sorted and grouped into deciles based on the ratio of the risk premium of the

Figure 9. Predictive model vs standard model



predictive model and the risk premium of the standard model. For each decile, the risk premium was calculated based on the actual values from the test data set. Graph 9 shows that the predictive risk model shows better predictive performance compared to the standard risk model. Namely, there is a greater correlation between the assessment of the risk premium using the predictive risk model and the actual risk premium compared to the standard risk model in each decile. It is clear from the graph that the predictive risk model more accurately predicts the actual risk premium for each decile, compared to the standard risk model.

5. CONCLUSION

The aim of this study was to examine the impact of risk classification, by using data mining methods on the non - life insurance ratemaking. For this purpose, development of a standard risk model, i.e. a standard approach in risk premium ratemaking based on the application of Poisson GLM for claims frequency estimate and Gamma GLM for claim severity estimate was investigated. As it was shown by the standard approach, GLM does not provide a reliable parameter estimates for MLF categorical predictors, in which individual categories have low risk exposure. Although GLMs are widely used for the purpose of determining insurance premiums, certain improvements to GLM, by using the data mining methods described in this paper, may solve practical problems in implementing the risk model for determining insurance premiums. The popularity of applying data mining methods in the actuarial community has been growing in recent years. The main reason for the increasing application of data mining methods in actuarial work comes from the efficiency and precision of these methods.

To improve GLM, CHAID and CR&T decision tree data mining methods were investigated. Both decision tree methods have proven to be very useful in selecting risk classes, as they easily group categories of MLF categorical predictors and require less time to prepare data, compared to the standard approach. The CHAID decision tree has better prediction performance compared to the CR&T decision tree, and is preferred because of the ease of application of the method. A predictive risk model was created, which combines risk classes using the CHAID decision tree with GLM. The results of comparing the risk premium estimated on the basis of the standard and predictive risk model with the actual risk premium showed that the predictive model more accurately estimates the risk premium, and thus has better forecasting performance.

The use of risk factor selection methods allows actuaries more time to refine the model, while reducing the risk that some of the important risk factors are not included in the model. In this study, only some of the data mining methods for risk classification are considered, and considering its importance, it opens up opportunities for further research. We believe it would be useful to investigate the results of applying data mining methods to risk classification on different data sets. Research conducted on a larger set of risk factors could also yield interesting results. In addition, the tested data mining methods can be easily applied to determining the premium of other types of non-life insurance.

Based on the available literature and the authors of best knowledge, the application of data mining methods in the non-life insurance premium pricing in the way presented in the paper, is presented for the first time in BiH and SEE region.

Appendix 1

Table A1. Variables' definitions

Variable	Explanation
MarkaD	dummy variable which codes vehicle brand (MarkaD=1=Skoda, MarkaD=2=Volkswagen, …MarkaD=29=Mitsubishi) based on exposure
KlasaD	dummy variable which codes vehicle class based on Euro car segmentation (KlasaD=1=C, KlasaD=2=D, ...KlasaD=7=M)
NamjenaD	dummy variable which codes policyholders who use vehicle for business purposes (NamjenaD=1) versus those who use vehicle for private purposes (NamjenaD=0)
LeasingD	dummy variable which codes policyholders who bought vehicle on leasing (LeasingD=1) versus those who don't (LeasingD=0)
Trajanje_ugD	dummy variable which codes policies with policy duration longer than 1 year (Trajanje_ugD=1) versus policies with annual duration (Trajanje_ugD=1)
Tip_osigD	dummy variable which codes policyholders with 6 and more vehicle (Tip_osigD=1) versus policyholders with up to 6 vehicles (Tip_osigD=0)

REFERENCES

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. and Thandi, N. 2007. A practitioner's guide to generalized linear models. casualty actuarial society. Towers Watson.

Antonio, K., and Valdez, E. A. 2010. Statistical concepts of a priori and a posteriori risk classification. Advances in Statistical Analysis 96 (2): 187-224.

Breiman, L., Friedman, J. H. , Olshen, R. A., and Stone, C. J. 1984. Classification and regression trees. New York: Chapman and Hall/CRC.

Brockman, M. J. and Wright, T. S. 1992. Statistical motor rating: making effective use of your data. Journal of the Institute of Actuaries 119: 457–543.

Coskun, S. 2016. Introducing credibility theory into GLMs for ratemaking on auto portfolio. Institute de Actuaries, Actuarial thesis. Centre d'Etudes Actuarielles.

Denuit, M., and Lang, S. 2004. Non-life rate-making with Bayesian GAMs. Insurance: Mathematics and Economics 35(3): 627-647.

Derrig, R. A. and Francis, L. 2006. Distinguishing the forest from the TREES: A Comparison of tree based data mining methods. Casualty Actuarial Society Forum.

Dionne, G. and Vanasse, C. 1988. A generalization of actuarial automobile insurance rating models: The negative binomial distribution with a regression component. ASTIN Bulletin 19 (2): 199-212.

Dionne, G. and Vanasse, C. 1992. Automobile insurance ratemaking in the presence of asymmetrical information. Journal of Applied Econometrics 7(2): 149-165.

Dugas, C., Bengio. Y., N. Chapados, N, Vincent, P., Denoncourt, G. and Fournier, C. 2003. Statistical learning algorithms applied to automobile insurance ratemaking. Casualty Actuarial Society Forum 1(1): 179-214.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. From data mining to knowledge discovery in databases. AI Magazine 17 (3): 37-54.

Flynn, M., and Francis, L. A. 2009. More flexible GLMs: zero-inflated models and hybrid models. Casualty Actuarial Society E-Forum, 148-224.

Francis, L. 2001. Neural networks demystified. Casualty Actuarial Society Forum, 253-320.

Frees, E. W. and Lee, G. 2016. Rating endorsements using generalized linear models casualty, Actuarial Society, Variance Advancing the Science of Risk 10(1): 51-74.

Garrido, J., Genest, C. and Schulz., J. 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics 70: 205-215.

Gallagher, C. A., Monroe, H. M. and Fish, J. L. 1990. An iterative approach to classification analysis. Casualty Actuarial Society, 237-281.

Goldburd, M., Khare, A. and Tevet, D. 2016. Generalized linear models for insurance rating. Casualty Actuarial Society, No. 5, 2nd edition.

Guo, L. 2003. Applying data mining techniques in property/casualty insurance. Casualty Actuarial Society Forum. Available at: https://www.casact.org/pubs/forum/03wforum/03wf001.pdf

Haberman, S. and Renshaw, A. E. 1996. Generalized linear models and actuarial science. The Statistician 45(4): 407-436.

Han, J., Kamber, M. and Pei, J. 2012. Data mining concepts and techniques (3rd ed.). Burlington, USA: The Morgan Kaufmann.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. The elements of statistical learning. New York, USA: Springer.

Hilbe, J. M. 2014. Modeling count data. New York: Cambridge University Press.

de Jong, P. and Heller, G. Z. 2013. Generalized linear models for insurance data (5th ed.). New York: Cambridge University Press.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society 29 (2): 119-127.

Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M. 2009. Modern Actuarial risk theory, using R. Berlin: Springer.

Kolyshkina, I., Wong, S. and Lim, S. 2004. Enhancing generalised linear models with data mining. casualty actuarial society. Discussion Paper Program.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models (2nd ed.). London: Chapman & Hall.

Meyers, G. 2007. Estimating loss costs at the address level. PowerPoint presentation at the CAS Predictive Modeling Seminar.

Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized linear models. Journal of the Royal Statistical Society 135(3): 370–384.

Ohlsson, E. and Johansson, B. 2010. Non-life insurance pricing with generalized linear models. Berlin: Springer-Verlag.

Parodi, P. 2014. Pricing in general insurance (1st ed.). New York: Chapman and Hall/CRC.

Renshaw, A. E. 1994. Modeling the claims process in the presence of covariates. ASTIN Bulletin 24 (2): 265-285.

SAS Institute. 2000. Data mining in the insurance industry - solving business problems using SAS enterprise miner software.

Shapiro, A. F. and Jain, L. C. 2003. Intelligent and other computational techniques in insurance. world scientific. https://doi.org/10.1142/5441

Sumathi, S. and Sivanandam, S. N. 2006. Introduction to data mining and its applications. Berlin: Springer-Verlag.

Tevet, D. 2013. Exploring model lift: is your model worth implementing?. Actuarial Review 40(2): 10-11.

Quinlan, J. R. 1986. Induction of decision trees. Machine Learning 1: 81-106.

Quinlan, J. R. 1993. C4.5 Programs for machine learning. Los Altos: Morgan Kaufmann.

Werner, G. and Modlin, C. 2010. Basic ratemaking (4th ed.). Casualty Actuarial Society.

Williams, B., Hansen, G., Baraban, A. and Santoni, A. 2015. A practical approach to variable selection - a comparison of various techniques. Casualty Actuarial Society E-Forum.

Yao, J. 2008. Clustering in ratemaking: with application in territories clustering. Casualty Actuarial Society Discussion Paper Program, 170–192.