Preliminary communication
UDC: 368.025.1:004.62
https://doi.org/10.18045/zbefri.2020.2.667

Risk factors selection with data mining methods for insurance premium ratemaking*

Amela Omerašević[1], *Jasmina Selimović*[2]

Abstract

Insurance companies that have adopted the application of data mining methods in their business have become more competitive in the insurance market. Data mining methods provides the insurance industry with numerous advantages: shorter data processing times, more sophisticated methods for more accurate data analysis, better decision-making, etc. Insurance companies use data mining methods for various purposes, from marketing campaigns to fraud prevention. The process of insurance premium pricing was one of the first applications of data mining methods in insurance industry. The application of the data mining method in this paper aims to improve the results in the process of non-life insurance premium ratemaking. The improvement is reflected in the choice of predictors or risk factors that have an impact on insurance premium rates. The following data mining methods for the selection of prediction variables were investigated: Forward Stepwise, Decision trees and Neural networks. Generalized linear models (GLM) were used for premium ratemaking, as the main statistical model for non-life insurance premium pricing today in most developed insurance markets in the world.

Key words: *GLM, data mining methods, forward stepwise, decision trees, neural networks*

JEL classification: *G22*

* Received: 30-11-2020; accepted: 22-12-2020

[1] *PhD Scholar, ABD, CFO, Uniqa osiguranje d.d. Sarajevo, Obala Kulina bana 19, 71000 Sarajevo, Bosnia and Herzegovina. Scientific affiliation: insurance, actuarial science. Phone: +387 33 289 000. Fax: +387 33 289 010. E-mail: amela.omerasevic@uniqa.ba.*

[2] *Associate professor, School of Economics and Business, University of Sarajevo, Trg oslobođenja – Alija Izetbegović 1, 71000 Sarajevo, Bosnia and Herzegovina. Scientific affiliation: insurance, actuarial science. Phone: +387 33 275 906. Fax: +387 33 275 966. E-mail: jasmina.selimovic@efsa.unsa.ba.*

1. Introduction

Insurance companies strive to strengthen their market position by introducing innovative solutions. The development of information technologies and the introduction of technological innovations, such as data warehousing, have enabled the development of data mining methods. Data mining is the process of discovering interesting patterns and knowledge from a large amount of data. Data mining uses a variety of methods to discover patterns and relationships between data to make valid predictions.

Data mining is a fairly new technology with great potential, which can help insurance companies focus on the most important information from the collected data in their business. Various data mining methods are used in the insurance industry, from classification, cluster analysis to regression. Data mining is used by insurance companies in various business areas, like insurance pricing, client acquisition, portfolio renewal, product development, fraud detection, reinsurance, marketing campaigns and claims assessment (SAS Institute, 2000). Insurance companies that have applied data mining methods have gained a competitive advantage in their operations. The main goal of this study was to examine the impact of the use of data mining methods for the risk factors selection on the non – life insurance premiums ratemaking. The research hypothesis states that adequate data mining methods used for risk factors selection of the non-life insurance premiums will improve the ratemaking process significantly in term of its quality.

In the paper, the impact of risk factors selected using data mining methods on the non-life insurance premium pricing is investigated. Risk factors are the characteristics of the insured, the subject of insurance and its environment that are believed to directly affect the insurance premium. Generalized linear models (GLM) as the standard for non-life insurance premium pricing in the European Union and in many other insurance markets were used for non-life insurance premium ratemaking. Data mining methods were used to select risk factors or predictors, which are included in the GLM to estimate insurance premiums.

The contribution of this paper is described through the application of the data mining method as a tool for the results' improvement in the process of non-life insurance premium ratemaking. The improvement is reflected in the choice of predictors or risk factors that have an impact on insurance premium rates.

This paper is organized in the following manner. Section 2 discusses the existing literature and gives some theoretical background. The elaborated literature covers the main findings of GLM in nonlife insurance. Also, first models for claims frequency and claims severity are presented. The data mining methods in insurance are widely used but regarding the topic of this paper authors presented data mining methods that can be used to overcome the problems of traditional GLMs, and to

improve performance of the risk premium predictive model. Section 3 provides information on data, model specification and applied methodology. The most commonly used approach in non-life insurance premium pricing is the risk premium approach. However, more accurate and flexible statistical models for premium ratemaking can be constructed by examining the claims frequency and claims severity separately. Out of GLM generally, separate GLM for claims frequency and GLM for claims severity are more efficient and practical to apply, so they have been used in this paper as well. Some of the most well-known data mining methods for variable selection were used to select risk factors in the paper: Stepwise regression, Decision trees and Neural networks. Section 4 presents empirical results and discussion. The research was conducted on motor hull insurance data set one of the leading insurance companies in Bosnia and Herzegovina. The data set contains 22 variables and consists 17,404 records on motor hull insurance policies during five consecutive years. Authors described the development and evaluation of claims frequency and claims severity models. Section 5 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 was introduced by Nelder and Wedderburn (1972). They show that GLM is an extension of traditional linear models, where the probability distribution of the dependent variable is a member of the family of exponential distributions (normal, Poisson, gamma, ...), and the expectation of the dependent variable is determined by a linear predictor based on nonlinear link function. The first application of GLMs in insurance premium pricing was the illustration of the application of GLM for the claim severity in motor insurance by McCullagh and Nelder (1989).

Most insurance companies in EU countries after the deregulation of the insurance market in the 1990s, took the opportunity to create their insurance tariffs based on GLM, as shown by some of the following papers. Renshaw (1994) showed how GLM can be used to analyze the claims frequency and claims severity based on individual data at the insured level. Brockman and Wright (1992) used GLIM software to statistically model the claims frequency and severity in premium pricing for motor liability insurance. Haberman and Renshaw (1996) presented a comprehensive overview of the application of GLM for various actuarial problems such as: survival models, multiple condition models, claims distribution models, insurance premium pricing and claims reserves in non-life insurance.

Over the years, many authors have made significant contributions to the development and improvement of GLM for non-life insurance premium pricing, and thus enabled the practical application of these models in the insurance

industry. A significant contribution belongs to Denuit and Charpentier (2005) who presented all aspects of insurance mathematics in a modern way. Anderson et al. (2007) is the most useful guide for actuaries for application of GLM in practice. The handbook, in addition to statistical theory and examples, provides insight into GLM construction – including data preparation and preliminary data analysis, model selection and number of iterations, model accuracy, and interpretation of model results. DeJong and Heller (2013) did an illustration of GLM for premium ratemaking on examples. Kaas et al. (2009) used GLM to determine premium motor third party liability insurance based on a bonus-malus system. Ohlsson and Johansson (2010) presented the basics of GLM theory for insurance tariff design and how to build a generalized linear model for calculating non-life insurance premium rates with example illustration, especially for multiplicative and hierarchical models. The paper presents useful extensions of GLM theory to Generalized Additive Models, with the application of interpolation of cubic and bivariate splines for modeling continuous independent variables. Goldburd et al. (2016) published a comprehensive manual for actuaries for the application of GLM in risk classification and tariff development for non-life types of insurance based on raw premium and claims data. Based on the idea of deJong and Heller (2013), Anderson et al. (2007) and Goldburd et al. (2016), we develop models for claims frequency and claims severity in this work.

In the last twenty years, data mining methods have become a useful tool in many areas of business. Industries are using data mining to achieve competitive advantage, increase efficiency, and provide better customer service (Fayyad et al. 1996). Data mining methods have also begun to be applied in the insurance industry. Sumathi and Sivanandam (2006) explored the concepts of data mining and data warehousing and presented the areas of application of data mining in the insurance industry. We used the work of Han et al. (2012) to get a basic knowledge of data mining methods. They presented a comprehensive overview of data mining methods, from classification and regression analysis to cluster analysis. Hastie et al., (2001) also systematically presented most of the statistical methods used today in data mining.

There are numerous papers on the application of data mining methods in the insurance industry. Some of the authors have attempted to use data mining methods for non-life insurance premium ratemaking. Lowe and Pryor (1996) compared the methods of neural networks and GLM and concluded that neural networks have a more general application than GLM and suggested certain possibilities of using neural networks in insurance but concluded that computationally demanding neural networks can prevent their wider application in insurance. Dugas et al. (2003) investigated the application of neural networks to motor insurance premiums ratemaking in North America. Guo (2003) described the application of the decision tree method to model claims frequency in non-life insurance. Yao (2008) used

cluster analysis methods to determine claims frequency by geographical areas. From the cited literature, it can be concluded that the inability to clearly present the results obtained on the basis of data mining models for insurance premium pricing is, at this moment, a problem which needs to be solved for the practical application of these methods in insurance companies and their acceptance by regulators.

At the same time, data mining methods can be used to overcome the problems of traditional GLMs, and to improve performance of the risk premium predictive model. Some of the authors combined data mining methods and GLM to take advantage of both approaches. The work of Kolyshkin et al. (2004) discusses the advantages of combining GLM with the Multivariate Adaptive Regression Splines (MARS) method. The results of this combined model were compared with the results of standard GLM. Williams et al. (2015) compared different data mining methods (Stepwise regression, decision tree, etc.) for the selection of predictors on the example of a household property insurance premium. Recent papers dedicated to GLM, by Makov and Weiss (2016), Coskun (2016) include Stepwise regression in variable selection. Makov and Weiss (2016), Guo (2003) used decision trees to select variables. Refaat (2007) proposes the application of a decision tree for variable selection, reduction of variable dimensionality and optimal discretization of continuous variables. Flynn and Francis (2009) used the CHAID decision tree in the data preparation process to estimate the claims frequency via GLM. Francis (2001) compared neural networks and regression models on insurance examples and used neural networks to select risk factors. The works of these authors were an incentive to explore data mining methods for the selection of variables in the data preparation phase, in order to improve the prediction performance and efficiency of the GLM risk premium predictive model.

3. Methodology

In the development of GLM for the insurance premium pricing, it is necessary to include all significant risk factors, i.e. prediction variables, which have an impact on the amount of insurance premium. Databases available to insurance companies, as well as available databases of external companies and institutions, contain hundreds of potential variables for the selection of risk factors. Manual selection of predictors i.e. risk factors takes a lot of time and therefore requires a large number of iterations during development phase of GLM. In the insurance industry, risk factors are in most cases defined on the basis of the existing insurance tariffs, research literature or statistical tests of the significance of prediction variables.

Data mining methods successfully address certain limitations of manual selection of predictors. The application of data mining methods aims to select significant predictors from the initial set of variables, which will be used to develop the

GLM for the claims frequency and claims severity. Data mining models are more accurate, faster, and more efficient in solving of business problems. The main reasons for the increasing attractiveness of data mining methods are:

- require less time required for data analysis,
- automatically selects the data to be used in pattern recognition,
- have the ability to process incomplete data and data with incorrect values,
- use test data sets to ensure the reliability of the results,
- provide a clear presentation of the results and useful feedback.

In order to improve the accuracy of risk premium predictions, this paper investigates data mining methods for the selection of risk factors or predictors.

3.1. Research methods

The most commonly used approach in non-life insurance premium pricing is the risk premium approach (Werner and Modlin, 2010). Risk premium is the average expected amount of claims under the insurance policy during the insurance period. In estimation of the risk premium, the distribution of total claims or the distribution of the claims frequency and the distribution of claim severity can be analyzed. Certain risk factors affect both the claims frequency and claims severity, while some risk factors affect only the claims severity or claims frequency. According to Klugman et al., (2004), more accurate and flexible statistical models for premium ratemaking can be constructed by examining the claims frequency and claims severity separately. Above all, separate GLM for claims frequency and GLM for claims severity are more efficient and practical to apply, so they have been used in this paper as well.

The application of data mining methods aims to select significant prediction variables from the initial set of variables, which will be used to develop GLM for claims frequency and GLM for claims severity. As the number of independent prediction variables that are the subject of research increases, there is a need to know the structure and relationships between these variables. The process of reducing the number of predictors under consideration is called data dimensionality reduction. For data dimensionality we use feature selection methods.

Feature selection (or variable selection) is the process of selecting a subset of significant predictors to be used to develop a model. Feature selection methods are useful due to:

- simplification of the model for easier interpretation,
- reducing the duration of model development,
- avoiding excessive parameterization of the model.

Some of the most well-known data mining methods for feature selection were used to select risk factors in the paper: Stepwise regression, Decision trees and Neural networks. The applied methods of data mining in the research are described in more detail below.

3.2. Generalized linear models (GLM)

Based on deJong and Heller (2013), we give only summary of the main characteristics of generalized linear models (GLMs). The purpose of GLM is to estimate the dependent response variable Y claims frequency and claims severity, based on a number of known independent predictors or risk factors X_i, where $i = 1, ..., n$. GLM is specified with following three components:

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

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

where parameter θ_i is related to the mean $\mu_i = E[Y_i]$, the scale parameter ϕ is fixed value estimated from data, while functions $a(.)$, $b(.)$ and $c(.)$ specifies a distribution function, suitable for solving GLM problems.

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

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

where: n is the number of data, p is a number of model parameters, $n - p$ degree of freedom and ξ_i offset. The parameters β_j are estimated by the method of maximum likelihood.

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

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

Link function $g(.)$ provides GLM flexibility in defining the relationship between mean and the linear predictor. The Log link function $g(x) = \ln(x)$ is most often used to determine the insurance premium, due to the ability to produce multiplicative models.

Advantages of using GLM over other methods of non-life insurance premium pricing are:

- GLM has a statistical framework, which provide techniques for standard error estimation, confidence interval, testing, model selection, and other statistical functions.
- Standard statistical software for calculating GLM makes the analysis of data for determining premium rates relatively easy (e.g. SAS, SPSS, R).
- GLM can be used separately to model the claims frequency and to model the claims severity with respect to the choice of independent variables.

3.3. Stepwise regression

Stepwise regression method selects the optimal number of predictors, by adding or removing independent predictors in each step, based on defined criteria. Although the stepwise regression method was introduced by Efroymson (1960), more than 60 years ago it's still a very popular data mining method, since it's computationally less demanding than searching for all possible combinations of predictors. The most common criticism directed at the application of the Stepwise regression method is based on the fact that standard statistical tests are not suitable for use in every step of the selection of prediction variables (Harrell, 2001). Despite criticism, this method has never ceased to be used and its use has revived for data mining purposes, in cases where a large number of potential predictors are present. Namely, Famoye and Rothe (2003) found that the application of stepwise regression, if taken with the appropriate degree of caution, is acceptable in practice.

There are several procedures for selecting variables using stepwise regression, most of which are applied: Forward selection, Backward elimination and Forward Stepwise regression. In this paper, Forward Stepwise Regression with AICC criteria is used for selection of risk factors. The process of Forward Stepwise regression of variables begins with the null model, i.e. without predictors in the model, and then in each step the least significant predictor is alternately excluded and the most significant predictor is included in the model. In each step, values are calculated that must be in accordance with the criterion variable.

Corrected Akaike Information Criterion (AICC) for entering or removing an effect from the current model use value:

$$AICC = Nln\left(\frac{(N-1)S_{yy} \times \tilde{r}_{yy}}{N}\right) + \frac{2p^r N}{N - p^r - 1}$$

where: N is number of data, p^r is number of parameters in the resulting model (including the intercept), SS_{yy} is weighted sample variance for y, $\tilde{r}_{yy}$ is the last diagonal element in the resulting $\tilde{R}$ matrix.

3.4. Decision trees

Decision tree method, as the name suggests, divides data by making a decision based on certain criteria. The speed and accuracy of algorithms, the ability to work with large data sets, without special requirements regarding the quality and relationship between variables, and the simplicity of presentation of results, are just some of the advantages of this method of data mining. Decision trees have the most common application for solving predictive problems with supervised learning. Decision trees according to the response variable are divided into classification decision trees with categorical response variable and regression decision trees with a continuous response variable. Building a decision tree starts from a single starting point called the root node, which contains the entire set of data at the top of the tree. The initial division is done using a prediction variable, dividing the data into two or more child nodes. A node that no more divisions, i.e. the one that reaches the end of the branch of the tree is called the final node or leaf. Data classification or prediction is done by going through a tree, starting from the roots to the leaves. The goal is to find a decision tree such that classification or prediction error is minimal.

The first decision tree algorithm known as Iterative Dichotomiser 3 (ID3) was developed by Quinlan (1986). Breiman et al. (1984) described the generation of binary Classification and Regression Trees (C&RT). Based on these algorithms, a numerous of decision tree induction algorithms have been developed: C4.5, QUEST, Random Forest, etc. Each of these algorithms has unique qualities in building a decision tree. Since CHAID and C&RT can be used to solve regression problems, they will be considered in this paper.

The C&RT tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, starting from the root node that contains the whole learning sample. For regression decision tree, C&RT use an impurity measure Least squares deviation – LSD, that is suitable for continuous response variables claims frequency and claims severity:

$$LSD(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} (y_i - \bar{y}(t))^2$$

where $N(t)$ is the number of observations at node t, y_i is the response value of observation i, $y(t)$ is the average response of the observations in t.

The average response is defined as:

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} y_i$$

CHAID (χ^2 automatic interaction detector) was developed by Kass (1980). CHAID decision tree development starts with identifying the target variable, and then splits the target into two or more child nodes using statistical algorithms. If dependent variable is continuous, the *F* test is used, while for categorical dependent variable the χ^2 test is used. For continuous response variables claims frequency and claims severity, the worth is based on the F test for the null hypothesis that the means of the response values are identical across the child nodes. The test statistic is

$$F = \frac{SS_{between}/(B-1)}{SS_{within}/(N(t)-1)}$$

where

$$SS_{between} = \sum_{b=1}^{B} N(t_b)(\bar{y}(t_b) - \bar{y}(t))^2, \quad SS_{within} = \sum_{b=1}^{B} \sum_{i=1}^{N(t_b)} (y_{bi} - \bar{y}(t_b))^2$$

where t_b denotes the b-th child node and B is the number of branches after splitting, $N(t)$ is the number of data in the decision tree and $N(t_b)$ is the number of data in node t_b. y_{bi} is the response variable for the value i in n-th node, $\bar{y}(t)$ is the average value of the response variable, while $y(t_b)$ is the average value of the response variable in node t_b.

3.5. Neural networks

Neural networks are one of the methods of artificial intelligence, which uses mathematical tools and the structure of the human brain when analyzing data. Neural networks have been successfully applied in medicine, education, banking, marketing, and other fields of social and technical sciences (Sumathi and Sivanandam, 2006). The development of neural networks began in the 1950s, when McCulloch and Pitts (1943) introduced the first model of an artificial neuron. Although the concept of artificial neural networks was introduced in the middle of the last century, they only became popular with the development of better performance databases and computers. Today, artificial neural networks are widely used in prediction, classification, optimization, image recognition, voice recognition and solving many other problems.

Numerous neural network algorithms have been developed, but the largest commercial use to date has been the error backpropagation algorithm. The backpropagation algorithm has the structure of a multilayer feedforward neural network and was first introduced by Rumelhart et al. (1986). The error backpropagation algorithm is a universal network learning algorithm applicable to prediction problems as well as to classification problems. The network structure consists of an input layer, one or more hidden layers, and an output layer. The error backpropagation algorithm starts at the input layer where the input data are presented. The inputs are weighted and received

by each node in the next layer. The weighted inputs are then summed and passed through the activation function to produce the output, which is weighted and passed to processing elements in the next layer. In this work error backpropagation algorithm is used with the network structure consists of an input layer, one hidden layers and an output layer. The neuron calculates its output based on the sigmoid activation function. For node j, this process is illustrated with equations:

$$I_j = \theta_j + \sum_{i=1}^{n} w_{ji} x_i \,,$$

$$y_j = f(I_j)$$

where: I_j is the activation level of node j, w_{ji} is the connection weight between nodes j and i, x_i is the input from node $i = 0,1, \ldots, n$, θ_j is the bias or threshold for node j, y_j the output of node j and $f(.)$ is the activation function.

4. Empirical data and results

The research was conducted on motor hull insurance data set one of the leading insurance companies in Bosnia and Herzegovina. Motor hull insurance paid compensation 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.

The data set contains 22 variables and consists 17,404 records on motor hull insurance policies during five consecutive years. The data set were divided by random distribution: 80% for training and 20% for testing and model evaluation. The insurance data of Motor Hull one of the leading insurance companies in BiH were used for the research. Having in mind that the Motor Hull insurance product has the same insurance coverage in all companies in BiH, and that the insurance company whose data were used in the survey has a significant share in the insurance market, we believe that the data used is a representative sample for this survey. The research was conducted with the CRISP-DM methodology using the IBM SPSS Modeller software package.

In this section, various data mining methods for selecting risk factors from the total number of prediction variables are discussed. Selected risk factors were used as input variables to estimate the claims frequency and the claims severity using GLM. The following data mining methods were used for risk factors selection:

- Forward Stepwise regression,
- CHAID decision tree,
- C&RT decision tree and
- Neural networks.

These data mining methods evaluate predictors by significance for the response variable. For decision trees (C&RT, CHAID) and neural networks, the importance of the prediction variable is calculated using sensitivity analysis. Sensitivity analysis measures how much the prediction error increases when one of the predictors is excluded. For more information on sensitivity analysis, see Saltelli et al. (2004), Francis (2001).

Models have been created for each of the feature selection methods, resulting in a smaller set of prediction variables. The methods were analyzed directly on the response variables: claims frequency or claims severity. All models were developed on a training data set, and model performance comparisons were conducted on a test data set. The obtained results are presented separately for the claims frequency and the claims severity.

In addition, a model with risk factors obtained using data mining methods was evaluated to determine whether data mining methods for the risk factors selection affect the improvement in the prediction of the claims frequency and the claims severity. The model assessment was performed by comparing the GLM for the claims frequency and the claims severity based on risk factors determined using the previously mentioned data mining methods. The assessment also includes a model that did not have a previous selection of variables using data mining methods, which we call the standard approach. Based on the model assessment, the best data mining method proposed for the risk factors selection. The results of the best ranking GLM for the claims frequency and the claims severity were used to develop a predictive risk model.

The criteria for ranking and selecting the best GLM with risk factors selected using data mining methods are:

1) Goodness of fit of the model was done on the basis of Akaike Information Criterion (AIC), developed by Akaike (1974). Information criteria represent the ratio between the accuracy of model adaptation to data and the complexity of the model. The lower the AIC information criterion, the better the model is considered.

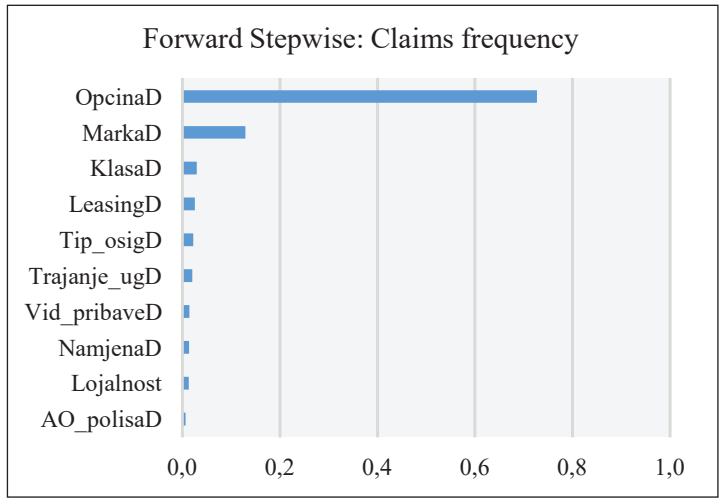
2) Predictive performance of the models was compared using the Gini coefficient (Meyers, 2007). The Gini coefficient, named after statistician and sociologist Corrado Gini, is commonly used in economics to measure national income inequality. The Gini coefficient does not quantify the profitability of a particular risk premium model but determines the model's ability to segment the best and worst risks. The higher the Gini coefficient, the better the predictive performance of the model.

4.1. Development and evaluation of claims frequency model

Risk factors selection for the claims frequency

The relative importance of each predictor in model assessment is shown using a graph of the significance of the prediction variables. The graphs of the significance of the prediction variables for the claims frequency are presented below, using each of the data mining method. Predictors relevant to the claims frequency response variable obtained from the Forward stepwise regression is shown in Figure 1.

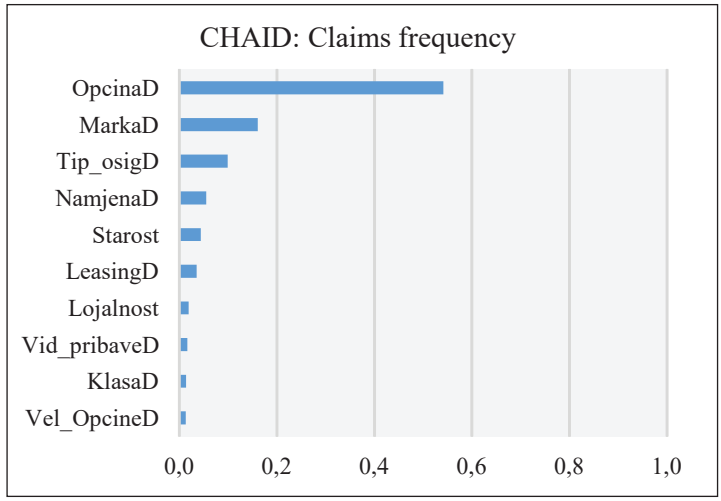
Figure 1: Predictors for claims frequency selected with Forward Stepwise



Source: Authors' calculations

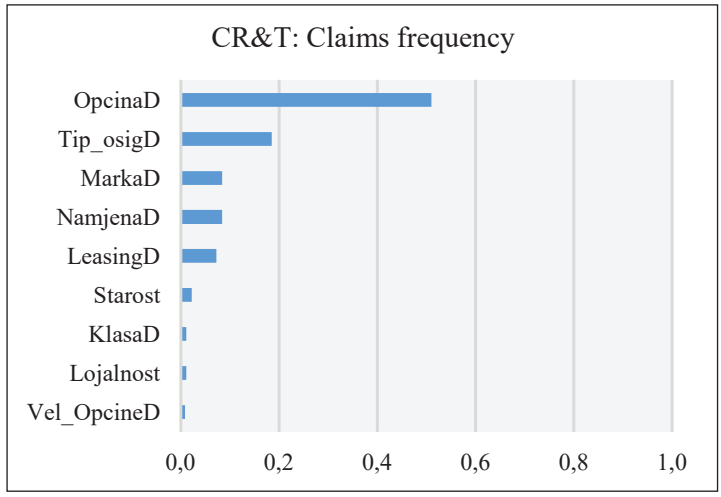
Prediction variables that affect the claims frequency of based on decision trees CHAID and CR&T are shown in Figure 2 and Figure 3. For both decision trees, no restrictions were used regarding the depth of the created tree, i.e. the number of tree levels. Tree growth is limited by defining a minimum number of nodes, in leaves 1%, and in the parent node 2%. The prediction results for both decision trees showed that the variable OpcinaD is the most significant prediction variable with over 50% significance, and the variables MarkaD, Tip_osigD and NamjenaD have the most influence on the claim frequency.

Figure 2: Predictors for claims frequency selected with CHAID



Source: Authors' calculations

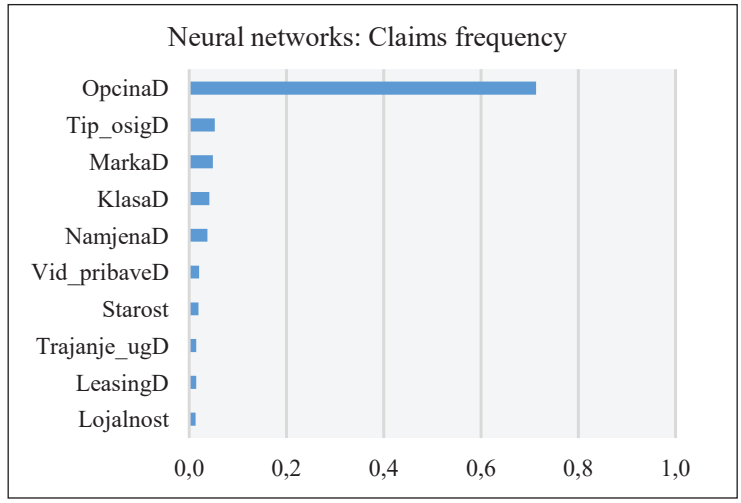
Figure 3: Predictors for claims frequency selected with CR&T



Source: Authors' calculations

Significant prediction variables for claims frequency selected, based on the neural network method are found in Figure 4. A backward error propagation algorithm was used to create the neural network model, with all prediction variables in input layer and response variable in output layer.

Figure 4: Predictors for claims frequency selected with Neural network



Source: Authors' calculations

The structure of the neural network consisted of only one hidden layer with seven neurons. Neural network prediction results show that OpcinaD with 71% is by far the most significant prediction variable for claims frequency.

GLM estimate of claims frequency

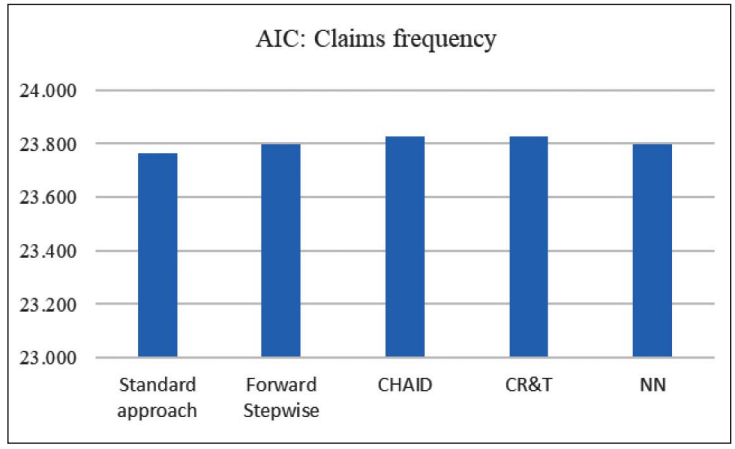
Significant predictors determined based on selected data mining methods are included in the claims frequency GLM. A Poisson GLM with a log link function was used to estimate the claims frequency. The Poisson distribution is the most common distribution for modeling the claims frequency according to Antonio and Valdez (2010), Dionne and Vanasse (1988, 1992), Denuit and Lang (2004), Flynn and Francis (2009). All insurance policies from the data set don't have the same risk exposure, *Log* (*Izloženosti*) is included in the model as an offset when calculating the number of claims. Statistically significant predictors using the Wald test and type III analysis were taken for risk factors. The Wald test follows a χ^2 distribution with a statistically significant value of $p \leq 0.001$.

A model evaluation was performed, i.e. a comparison GLMs with predictors selected on the basis of the previously mentioned data mining methods and GLM standard approach. In standard approach all prediction variables are included in the GLM to get significant predictors.

The AIC information criterion of the GLM Poisson model for claims frequency does not show significant differences between different risk factor selection

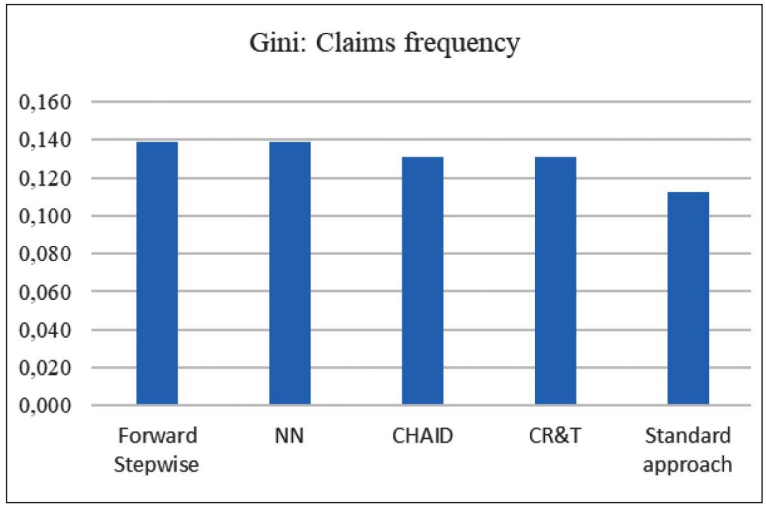
methods (Figure 5). The reason lies in the fact that all methods extract a similar number of risk factors from the total number of prediction variables. The Forward Stepwise regression and Neural networks show a better result compared to other methods.

Figure 5: AIC – Claims frequency model ranking



Source: Authors' calculations

Figure 6: Gini – Claims frequency model ranking



Source: Authors' calculations

Analysis the prediction performance of the Poisson GLM based on Gini coefficients (Figure 6), shown following conclusions:

1) Forward stepwise regression and Neural networks methods achieve very good results in claims frequency estimate.
1) Decision tree methods CHAID and CR&T achieve equal but slightly lower results for the claims frequency estimate.
2) Application of any of the data mining methods in the selection of risk factors for Poisson GLM, achieves better results for the claims frequency compared to the standard approach.

From the above, it can be concluded that the selection of the optimal number of prediction variables using Forward stepwise regression and Neural networks before inclusion in GLM improves the predictive performance of the model for claims frequency. Given the simplicity of application and interpretation of results, as well as the speed of execution of Forward stepwise regression has an advantage in the risk factors selection in comparison with Neural networks. The better predictive performance of the GLM Poisson model with selected risk factors using Forward stepwise regression resulted with the inclusion of additional significant predictor OpcinaD compared to the standard approach (Table 1). The selected predictors are statistically significant and have an impact on the claims frequency. Choosing the optimal number of risk factors using Forward stepwise regression before inclusion in GLM improves the predictive performance of the claims frequency estimate.

Table 1: Poisson GLM: Risk factors

Prediction variables	Standard approach			Forward Stepwise		
	Wald χ^2	df	p-value	Wald χ^2	df	p-value
Intercept	34.149	1	0.000	163.968	1	0.000
MarkaD	21.543	3	0.000	88.842	28	0.000
KlasaD	56.877	3	0.000	45.570	6	0.000
NamjenaD	22.559	1	0.000	19.384	1	0.000
LeasingD	45.687	1	0.000	17.578	1	0.000
Trajanje_ugD	30.758	1	0.000	26.907	1	0.000
Tip_osigD	33.209	1	0.000	20.893	1	0.000
OpcinaD				540.358	91	0.000

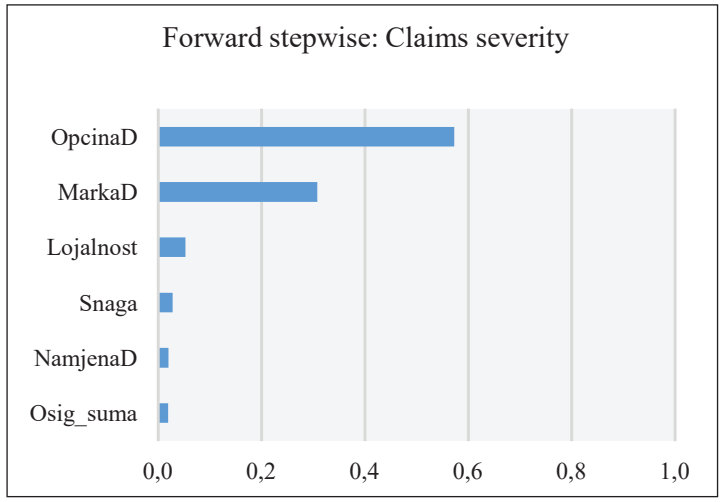
Source: Authors' calculations

4.2. Development and evaluation of a model for claims severity

Risk factors selection for claims severity

The results of the significance of the prediction variables for claims severity, using each of the data mining methods, are presented below using a significance graph. Significant prediction variables based on the Forward stepwise regression for claims severity are shown in Figure 7.

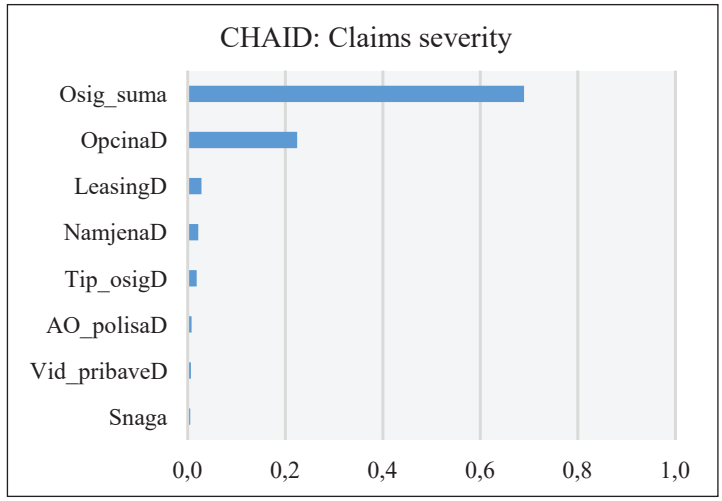
Figure 7: Predictors for claims severity selected with Forward Stepwise



Source: Authors' calculations

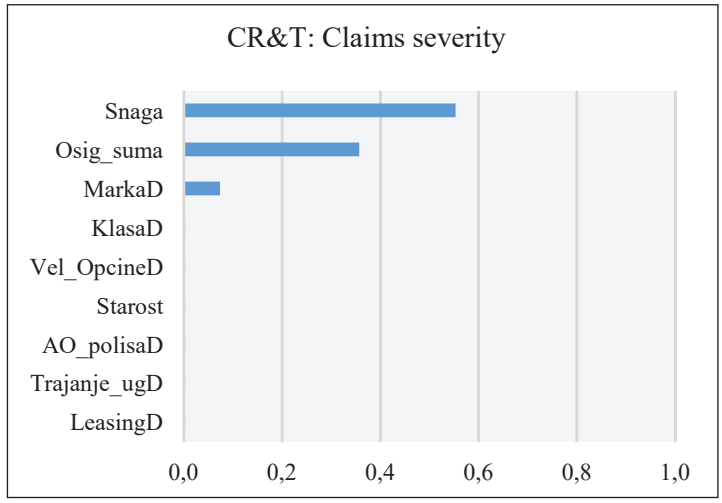
Figure 8 and Figure 9 show predictors for claims severity selected, based on the CHAID and CR&T decision tree. For the growth constraints of both decision trees, only a minimum number of nodes was defined, in leaves 1%, and in the parent node 2%. The CHAID decision tree select two variables Osig_suma and OpcinaD with over 90% impact on average claim. The CR&T decision tree select Osig_suma and SnagaD with over 90% significance for the average claim.

Figure 8: Predictors for claims frequency selected with CHAID



Source: Authors' calculations

Figure 9: Predictors for claims frequency selected with CR&T

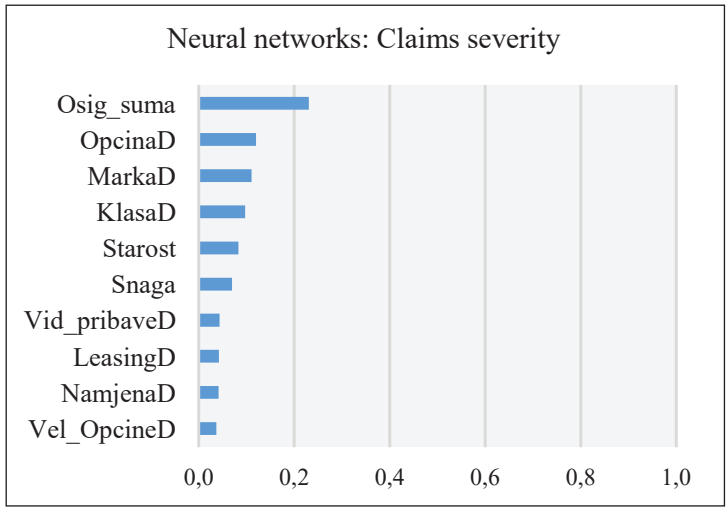


Source: Authors' calculations

Figure 10 shows significant predictors for claims severity selected, based on the Neural network method. The neural network model was created using a backpropagation error algorithm with all original prediction variables in the input layer, seven neurons in one hidden layer, and a claims severity in the output

layer. The results of the prediction via the Neural network show that Osig_suma, OpcinaD and MarkaD and KlasaD are the most significant predictors selected for claims severity.

Figure 10: Predictors for claims severity selected with Neural network



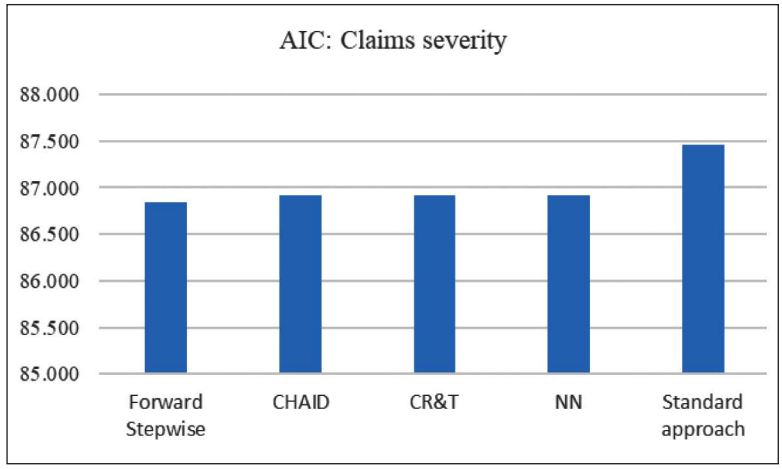
Source: Authors' calculations

GLM estimate for claims severity

Based on the data mining methods for variable selection, the optimal number of variables included in Gamma GLM with a log link function to estimate the claims severity was obtained. The Gamma distribution was used to model the claims severity as the most popular distribution for modeling the claims severity based on actuarial literature (Ohlsson and Johansson, 2010; Parodi, 2014; Kaas et al., 2009). To achieve the multiplicative model, instead of the canonical link function, the log link function was used. For scale parameter ϕ, Pearson's moment estimator was used. Most predictors are not statistically significant and have no effect on the claim severity. All variables with a p-value greater than 0.001 according to the Wald test were excluded from the model.

The AIC shows very similar results for different gamma GLM models to estimate the claims severity. GLM for the claims severity with risk factors selected based on the Forward stepwise regression and CHAID decision tree show a better result compared to other models. The ranking of methods based on AIC criteria is shown in Figure 11.

Figure 11: AIC – Claim severity model ranking

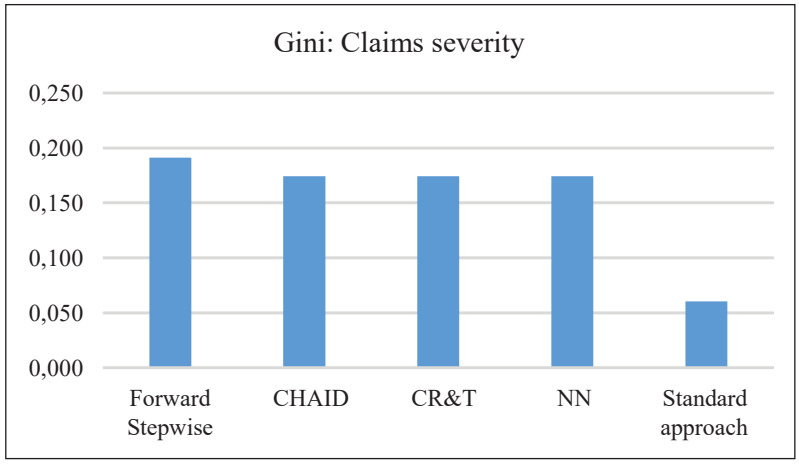


Source: Authors' calculations

The prediction performance of the Gamma GLM model was analyzed using Gini coefficients (Figure 12). Based on the Gini coefficients the following conclusions can be made:

1) Models with risk factors selected with Forward stepwise regression and CHAID decision tree achieve better results for the claims severity, compared to other methods.
2) Models with risk factors obtained using the CR&T method and the Neural Network achieve the same, but slightly worse results.
3) Application of any of the data mining methods in the selection of risk factors for Gamma GLM, achieves better results for the claims severity compared to the standard approach.

Figure 12: Gini – Claims severity model ranking



Source: Authors' calculations

From the above, it can be concluded that the selection of the optimal number of prediction variables using the methods of Forward stepwise regression and CHAID decision tree, before the inclusion of GLM, improves the predictive performance of the claims severity model. The results obtained based on standard aproach for gamma GLM suggest that the average amount of damage to the analyzed portfolio is affected by the Loyalty of the insured (Table 2). Using the Forward stepwise regression method, three risk factors were identified, which are statistically significant for the response variable average amount of damage. Forward stepwise regression identified more risk factors than standard approach and other data mining methods and as result give more precisely assessment. The risk factor that affect the claims severity differs significantly from the risk factors that affect the claims frequency.

Table 2: Gamma GLM: Risk factors

Prediction variables	Standard approach			Forward Stepwise		
	Wald χ^2	df	p-value	Wald χ^2	df	p-value
(Intercept)	81,665.391	1	0.000	18,487.290	1	0.000
Lojalnost	20.254	2	0.000	24.436	2	0.000
NamjenaD				12.498	1	0.000
Osig_suma				178.763	1	0.000

Source: Authors' calculations

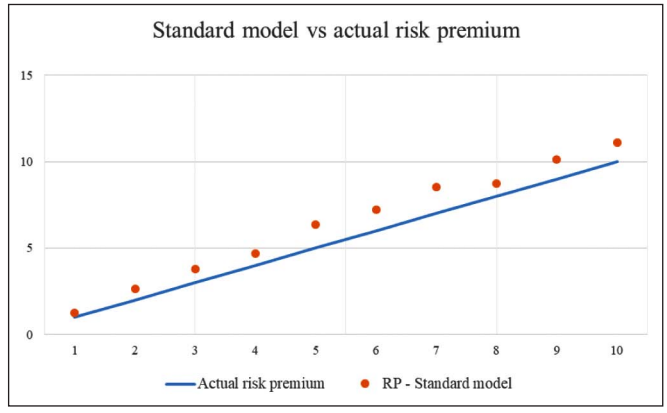
4.3. Model evaluation

From the evaluation of claims frequency and claims severity models we can conclude that Forward stepwise regression showed excellent performance and this method is the best candidate for the predictive risk model. Given the ease of application and the speed of development of Forward stepwise regression model, the time required to select risk factors can be reduced, using these methods. The based on GLM for claims frequency and claims severity predictive and standard risk models are created. The risk premium is obtained by multiplying the relativity of the claims frequency and the relativity of the claims severity.

For the predictive risk model, the GLM model for the claims frequency and the claims severity for which risk factors were selected using Forward stepwise regression were chosen. As a standard risk model, GLM was used for the claims frequency and claims severity with predictive variables that were not previously selected using data mining methods.

An evaluation of the predictive risk model in relation with to the standard risk model was performed. The simplest way to compare the predictive and standard risk models is to graphically compare the performance of the estimated risk premiums of both models, relative to the actual risk premium. Figure 13 shows a comparison of the estimated risk premium of the standard risk model in relation to the actual risk premium, on test data grouped in deciles.

Figure 13: Standard model risk premium vs actual risk premium

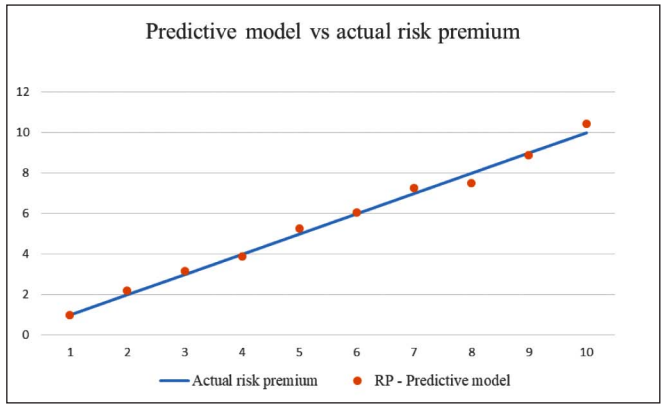


Source: Authors' calculations

Figure 14 shows a comparison of the estimated risk premium of the predictive risk model in relation to the actual risk premium on test data grouped in deciles. Prior to grouping the data, the records were sorted according to the value of the

estimated risk premiums. For each group, the average values of estimated and actual risk premiums were calculated, and the log of these values is shown in the graph. Both graphs clearly show that the estimated values of the risk premium of the predictive model have a better adjustment to the actual risk premiums compared to the standard model. Although both models meet the set business objectives, we can conclude that the predictive risk model shows better predictive performance compared to the standard model.

Figure 14: Predictive vs actual risk premium



Source: Authors' calculations

5. Results and discussion

From the evaluation of claims frequency and claims severity models we can conclude that both methods of Forward stepwise regression showed excellent performance, and are the best candidate for the predictive risk model. Above all, GLM models will be more accurate and will achieve better assessment results. Both methods select the same risk factors for the claims frequency and claims severity. It can be expected that these methods will show certain differences in the results on a larger number of prediction variables.

From the previous, we can conclude that premium rates are strongly influenced by selected risk factors. A better understanding of the real impact of all risk factors on the claims frequency and the claims severity can help insurance companies to offer adequate premium to policyholder. The adequate premium is of the utmost importance for insurance company every day operations and for the economy as well. Namely, if the calculated premium is not adequate (too low), the technical provision consequently will be lower than needed, and it is not likely that in the

case of damage it will cover the claim (and therefore the substance of insurance is ruined). On the other hand, if the calculated premium is too high, the technical provisions will be high influencing the financial result of the company, but economically wise, too high premium is not suitable for the market.

If an insurance company does not have an adequate premium ratemaking, 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 competition premium, which will lead to financial loss of insurance companies. Data mining methods are more efficient and comprehensive in the selection of risk factors in determining adequate non-life insurance premium rates compared to the standard approach.

Regarding the contribution to the scientific field, 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.

Beside the motor hull the results of this research are representative and can be applied to the motor third party liability insurance, household insurance, travel health insurance, accident insurance and other types of non-life insurance with homogeneous groups of policyholders. The research is not applicable for individual risk premiums ratemaking, such as insurance of corporate clients.

6. Conclusion

In the process of developing a predictive model for non-life insurance premium pricing, data mining methods were used for selection of risk factors that have an impact on the insurance premium. For the risk factors selection in the study, the following data mining methods were considered: Forward stepwise regression, CHAID decision tree, C&RT decision tree and neural networks. Data mining methods for variable selection aim to find the best possible combination of predictors by reducing dimensionality, which contributes to simplifying the interpretation of model results, shortening model development time, and avoiding excessive model parameterization. Models were created for each of the variable selection methods, which resulted in a smaller set of significant prediction variables. The methods were analyzed directly on the response variables: claims frequency and claims severity. Selected risk factors with data mining methods were included in the Poisson GLM for the claims frequency and the Gamma GLM for claims severity estimate. Goodness of fit of the models and predictive performance of the models were performed.

Very good results for the risk factors selection were achieved by the Forward stepwise regression, which is also the easiest to implement. Forward stepwise

regression method has identified a number of risk factors for both the claims frequency and the claims severity and thus improved the predictive performance of the GLM compared to the standard approach when method for risk factors selection wasn't use before inclusion in GLM. Other data mining methods used to select risk factors have shown satisfactory results in terms of improving forecast performance, compared to the standard approach. 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.

The conclusions of this research are representative and can be applied other non-life insurance types. The research is not applicable for individual risk premiums ratemaking, such as insurance of corporate clients.

In this study, application some of data mining methods in insurance pricing are considered, which opens up opportunities for further research. It would be useful to investigate the results of applying data mining methods on the larger data set of risk factors or applied research on some other types of non-life insurance. We consider these to be some of the useful areas for further research.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions*, Vol. 19, No. 6, pp. 716–723, https://ieeexplore.ieee.org/document/1100705.

Anderson, D. et al. (2007) *A Practitioner's Guide to Generalized Linear Models*, CAS, https://www.aktuarai.lt/wp-content/uploads/2018/06/Anderson_et_al_Edition_3.pdf.

Antonio, K., Valdez, E. A. (2010) 'Statistical concepts of a priori and a posteriori risk classification', *Advances in Statistical Analysis*, Vol. 96, No. 2, pp. 187–224.

Breiman, L. et al. (1984) *Classification and Regression Trees*, New York: Chapman and Hall/CRC.

Brockman, M.J., Wright, T.S. (1992) 'Statistical motor rating: making effective use of your data', *Journal of the Institute of Actuaries*, Vol. 119, pp. 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., Charpentier, A. (2005) Mathematiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement, *Economica.*

Denuit, M., Lang, S. (2004) 'Non-life rate-making with Bayesian GAMs'. *Insurance: Mathematics and Economics*, Vol. 35, No. 3, pp. 627–647, https://www.sciencedirect.com/science/article/abs/pii/S0167668704000940.

Dionne, G., Vanasse, C. (1988) 'A generalization of actuarial automobile insurance rating models: The negative binomial distribution with a regression component', *ASTIN Bulletin*, Vol. 19, No. 2, https://ideas.repec.org/p/mtl/montec/8833.html.

Dionne, G., Vanasse, C. (1992) 'Automobile Insurance Ratemaking in the Presence of Asymmetrical Information', *Journal of Applied Econometrics,* Vol. 7, No. 2, pp. 149–165, https://www.researchgate.net/publication/5140149_Automobile_Insurance_Ratemakin g_in_the_Presence_of_Asymmetrical_Information.

Dugas, C. et al. (2003) 'Statistical learning algorithms applied to automobile insurance ratemaking', *Casualty Actuarial Society Forum*, Vol. 1, No. 1, pp. 179–214, https://www.researchgate.net/publication/247643985_Statistical_Learning_Algorithms _Applied_to_Automobile_Insurance_Ratemaking.

Efroymson, M.A. (1960) 'Multiple Regression Analysis. In: Mathematical Methods for Digital Computers', *New York: John Wiley & Sons, Inc.*

Famoye, F. & Rothe, D.E. (2003) 'Variable Selection for Poisson Regression Model', *Journal of Modern Applied Statistical Methods*, Vol. 2, No. 2, pp. 380–388, https://core.ac.uk/download/pdf/195154175.pdf.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine,* Vol. 17*,* No. 3, https://ojs.aaai.org//index.php/aimagazine/article/view/1230.

Flynn, M., Francis, L. A. (2009) *More Flexible GLMs*: *Zero-Inflated Models* and *Hybrid Models,* Casualty Actuarial Society E-Forum, pp. 148–224, https://www.semanticscholar.org/paper/More-Flexible-GLMs-Zero-Inflated-Models-and-Hybrid-Flynn-Francis/0d195fec002e53f87255fef59c9822ca184997d1.

Francis, L. (2001) 'Neural Networks Demystified', *Casualty Actuarial Society Forum*, pp. 253–320, https://www.casact.org/pubs/forum/01wforum/01wf253.pdf.

Goldburd, M., Khare, A., Tevet, D. (2016) 'Generalized Linear Models for Insurance Rating', CAS, https://www.casact.org/pubs/monographs/papers/05-Goldburd-Khare-Tevet.pdf.

Guo, L. (2003) 'Applying data mining techniques in property/casualty insurance', CAS Forum, https://www.casact.org/pubs/forum/03wforum/03wf001.pdf.

Haberman, S., Renshaw, A.E. (1996) 'Generalized linear models and actuarial science', *The Statistician*, Vol. 45, No. 4, pp. 407–436, https://www.researchgate.net/publication/271865941_Generalized_Linear_Models_and_Actuarial_Science.

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

Harrell, F.E. (2001) *Regression modeling strategies: with applications to linear models, logistic regression and survival analysis*, New York: Springer.

Hastie, T., Tibshirani, R., Friedman, J. (2001) *The Elements of Statistical Learning*, New York, USA: Springer.

deJong, P., 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, Series C, Vol.* 29, No. 2), pp. 119–127, https://www.jstor.org/stable/2986296?seq=1.

Kaas, R. et al. (2009) *Modern Actuarial Risk Theory*, using R. Berlin: Springer.

Klugman, S.A., Panjer, H.H., Wilmot, G.E. (2004) *Loss Models, from Data to Decisions*, 2nd ed., New York, SAD: Wiley.

Kolyshkina, I., Wong, S., Lim, S. (2004) 'Enhancing Generalised Linear Models with Data Mining', Casualty Actuarial Society, Discussion Paper Program, https://www.casact.org/pubs/dpp/dpp04/04dpp279.pdf.

Lowe, J., Pryor, L. (1996) 'Neural networks v. GLMs in pricing general insurance', *General Insurance Convention,* https://www.actuaries.org.uk/system/files/documents/pdf/0417-0438.pdf.

Makov, U., Weiss, J. (2016) 'Predictive Modeling for Usage-Based Auto Insurance', In E. Frees, G. Meyers, & R. Derrig (Eds.), *Predictive Modeling Applications in Actuarial Science* (International Series on Actuarial Science, pp. 290–308), Cambridge: Cambridge University Press, https://www.cambridge.org/core/books/predictive-modeling-applications-in-actuarial-science/predictive-modeling-for-usagebased-auto-insurance/E4693E09E8FCD6A86885ABE81270B1C3.

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

McCulloch, W.S., Pitts, W. (1943) 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of mathematical biophysics*, No. 5, pp. 115–133, https://link.springer.com/article/10.1007/BF02478259.

Meyers, G., (2007) 'Estimating Loss Costs at the Address Level', PowerPoint presentation at the CAS Predictive Modeling Seminar.

Nelder, J. A., Wedderburn, R. W. M. (1972) 'Generalized linear models', *Journal of the Royal Statistical Society, Series A,* Vol. 135, No. 3, pp. 370–384, https://www.jstor.org/stable/2344614?seq=1.

Ohlsson, E., 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.

Refaat, M. (2007) *Data Preparation for Data Mining Using SAS*, Morgan Kaufmann Publishers.

Renshaw, A. E. (1994) 'Modeling the Claims Process in the Presence of Covariates', *ASTIN Bulletin*, Vol. 24, No. 2 pp. 265–285, https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/modelling-the-claims-process-in-the-presence-of-covariates/144A9480B73604F1F456E8C80D27F74D.

Rumelhart, D., Hinton, G., Williams, R. (1986) 'Learning representations by back-propagating errors', *Nature*, No. 323, pp. 533–536, https://www.nature.com/articles/323533a0.

Saltelli, A., S. et al. (2004) *Sensitivity Analysis in Practice – A Guide to Assessing Scientific Models*, John Wiley.

SAS Institute (2000) *Data Mining in the Insurance Industry – Solving Business Problems Using SAS Enterprise Miner Software.* A SAS White Paper.

Sumathi, S., Sivanandam, S.N. (2006) *Introduction to Data Mining and its Applications*, Berlin: Springer-Verlag.

Quinlan, J.R. (1986) 'Induction of decision trees', *Machine Learning*, No. 1, pp. 81–106, https://link.springer.com/article/10.1007/BF00116251.

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

Williams, B. et al. (2015) *A Practical Approach to Variable Selection – A Comparison of Various Techniques*, Casualty Actuarial Society E-Forum, https://www.casact.org/pubs/forum/15sumforum/Williams-Hansen-Baraban-Santoni.pdf.

Yao, J. (2008) 'Clustering in ratemaking: with application in territories clustering', *Casualty Actuarial Society Discussion Paper Program*, pp. 170–192, https://www.researchgate.net/publication/241526333_Clustering_in_Ratemaking_with_Application_in_Territories_Clustering.

Odabir faktora rizika metodama rudarenja podataka za izračun stopa premije osiguranja

Amela Omerašević[1], *Jasmina Selimović*[2]

Sažetak

Osiguravajuća društva koja su prva usvojila primjenu metoda rudarenja podataka u svom poslovanju postali su konkurentniji na tržištu osiguranja. Metode rudarenja podataka osiguravajućoj industriji pružaju brojne prednosti: kraće vrijeme obrade podataka, sofisticiranije metode za precizniju analizu podataka, bolje donošenje odluka itd. Osiguravajuća društva koriste metode rudarenja podataka u razne svrhe, od marketinških kampanja do sprečavanja prijevara, a među prvima je ta metoda bila u postupku određivanja premija osiguranja. Primjena metode rudarenja podataka u ovom radu ima za cilj poboljšati rezultate u procesu izračuna stope premije neživotnih osiguranja. Poboljšanje se ogleda u odabiru varijabli predviđanja ili faktora rizika koji utječu na stope premija osiguranja. Istražene su sljedeće metode rudarenja podataka za odabir varijabli predviđanja: Postepena regresija, Stabla odlučivanja i Neuronske mreže. Za izračun premijskih stopa korišteni su Generalizirani linearni modeli (GLM), koji su danas glavni statistički model određivanja premija neživotnih osiguranja u većini razvijenih tržišta osiguranja u svijetu.

Ključne riječi: *GLM, metode rudarenja podataka, postepena regresija, stabla odlučivanja, neuronske mreže*

JEL klasifikacija: *G22*

[1] *Doktorandica, CFO, Uniqa osiguranje d.d. Sarajevo, Obala Kulina bana 19, 71000 Sarajevo, Bosna i Hercegovina. Znanstveni interes: osiguranje, aktuarstvo. Tel.: +387 33 289 000. Fax: +387 33 289 010. E-mail: amela.omerasevic@uniqa.ba.*

[2] *Izvanredna profesorica, Ekonomski fakultet, Univerzitet u Sarajevu, Trg oslobođenja – Alija Izetbegović 1, 71000 Sarajevo, Bosna i Hercegovina. Znanstveni interes: osiguranje, aktuarstvo. Tel.: +387 33 275 906. Fax: +387 33 275 966. E-mail: jasmina.selimovic@efsa.unsa.ba.*