Aktüerya Derneği

İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 16, 2023, 2, 81-99

Geliş / Received:04.12.2023, **Kabul** / Accepted: 28.12.2023

Araştırma Makalesi / Research Article

Regression Tree Approach to Estimation of Health Insurance Premium

Başak Bulut Karageyik
Hacettepe University
Department of Actuarial Sciences, Ankara, Türkiye
basakbulut@hacettepe.edu.tr
ORCID:0000-0003-4080-9165

Abstract

This paper proposes an approach to predicting insurance premiums in health insurance by combining traditional generalized linear models (GLM) with advanced machine learning-driven regression tree analysis. The study first uses GLM on real complementary health insurance data to examine the importance of variables, focusing on those variables that have a large impact on premium estimates. Subsequently, it is investigated whether the variables identified as significant by GLM can also be identified as significant by regression tree analysis. In the application of machine learning, the effect of stratified sampling in accordance with the data structure in terms of the risk variables considered in premium forecasts is also analyzed. This study contributes to the actuarial understanding of premium estimation and provides insurers with a concrete framework to help them negotiate the complex world of health insurance data. By integrating the advantages of GLM and regression trees, this study provides a comprehensive comparison for insurers to adapt to changing risk factors. This study represents a innovative attempt to incorporate a regression tree methodology, providing a novel and accurate estimation of premium amounts in the realm of insurance analysis.

Keywords: Actuarial premium estimation, Regression tree, Machine learning techniques, Generalized linear models,

Öz

Sağlık Sigortası Primi Tahmininde Regresyon Ağacı Yaklaşımı

Bu çalışma, geleneksel genelleştirilmiş doğrusal modelleri (GLM) gelişmiş makine öğrenimi odaklı regresyon ağacı analizi ile birleştirerek sağlık sigortasında sigorta primlerini tahmin etmeye yönelik bir yaklaşım önermektedir. Çalışmada ilk olarak değişkenlerin önemini incelemek için gerçek tamamlayıcı sağlık sigortası verileri üzerine GLM uygulanmakta ve prim tahminleri üzerinde büyük etkisi olan değişkenlere odaklanılmaktadır. Daha sonra, GLM tarafından önemli olarak tanımlanan değişkenlerin regresyon ağacı analizi ile de önemli olarak tanımlanıp tanımlanamayacağı araştırılmaktadır. Makine öğrenmesi uygulamasında, prim tahminlerinde dikkate alınan risk değişkenleri açısından veri yapısına uygun olarak tabakalı örneklemenin etkisi de analiz edilmektedir. Bu çalışma, prim tahminine ilişkin aktüeryal anlayışa katkıda bulunmakta ve sigortacılara sağlık sigortası verilerinin karmaşık dünyasında müzakere etmelerine yardımcı olacak somut bir çerçeve sunmaktadır. GLM ve regresyon ağaçlarının avantajlarını bir araya getiren bu çalışma, sigortacıların değişen risk faktörlerine uyum sağlamaları için kapsamlı bir karşılaştırma sunmakta ve sigorta analizi alanında prim tutarlarının yeni ve doğru bir şekilde tahmin edilmesini sağlayan bir regresyon ağacı metodolojisini içeren yenilikçi bir çalışmayı temsil etmektedir.

Anahtar sözcükler: *Aktüeryal prim tahmini, Regresyon ağacı, Makine öğrenme teknikleri, Genelleştirilmiş doğrusal modeller*

1. Introduction

Actuarial science brings essential insights with a statistical, demographic, and social perspective into the analysis of risk factors that influence complicated insurance occurrences. Hence, an actuarial perspective enhances the ability to manage the complex world of risk. In the ever-evolving landscape of actuarial science, the estimation of insurance premiums stands as intricate with mathematical precision, statistical insight, and an acute understanding of risk dynamics. The premium estimation is the most important and lies at the heart of insurance pricing, financial sustainability, and risk management strategies. The estimation of insurance premiums holds paramount importance within the domain of actuarial science and the broader insurance industry.

The significance of premium estimation is multifaceted, encompassing financial stability, risk management, market competitiveness, and the overall sustainability of insurance operations. Premium estimation in actuarial science involves the use of various techniques and models to assess risk and determine the appropriate pricing for insurance coverage. In premium estimation techniques, as a combination of traditional and modern methodologies, frequency and severity models, generalized linear models (GLM), the loss ratio method, credibility theory, Bayesian methods, time series analysis, and extreme value theory (EVT) are commonly used in actual science. Among these methods, GLMs is one of the most preferred because it extends traditional linear models to handle non-normally distributed response variables. Actuaries use GLMs to model relationships between premiums and risk factors, incorporating link functions that account for the specific distribution of the response variable (e.g., Poisson or Gamma distributions).

GLMs play a pivotal role, especially in non-life insurance, in assessing and pricing risks, as well as in estimating more accurate reserves. Actuarial science often involves the application of statistical models, including GLMs, for analyzing and modeling insurance-related data. The GLM is developed as actuarial illustrations in the standard text by McCullagh and Nelder [1]. They provide numerous instances of how GLMs have been fitted to other kinds of data, such as average claim costs from a portfolio of auto insurance. Then Renshaw [2] and Renshaw and Verall [3] made the first studies in the actuarial field. In 1996, Haberman and Renshaw [4] analyzed in detail the use of GLMs in actuarial data analysis and demonstrated the use of GLMs in insurance claim frequency and severity. Over the years, GLM has been

applied in the calculation of loss resezrves, credibility, and mortality forecasting. These references cover the theoretical foundations as well as practical applications of GLMs in the actuarial sciences: Dobson [5], Anderson et al. [6], Antonio and Beirlant [7], De Jong and Heller [8], Wüthrich and Merz [9], Ohlsson et al. [10], and Frees [11].

A decision tree is a graphical representation and predictive modeling tool used in machine learning and data analysis. Decision trees are particularly useful for classification and regression tasks, as they help break down complex decision-making processes into a series of simpler, interpretable steps based on the input features of the data. CART is a versatile type of decision tree that can be used for both classification and regression tasks. It recursively splits the dataset based on the most significant attribute at each node. According to its purpose, CART is divided into two parts: firstly, to classify the data into discrete classes or categories, and secondly, to predict numerical values, making it suitable for the regression task.

Regression trees are an important instrument in the actuarial toolbox since they offer a special perspective for understanding and evaluating intricate risk dynamics. Regression trees excel at identifying distinct segments within a dataset, enabling actuaries to tailor risk assessment strategies to specific groups. Quan [12] summarized the advantages of the tree-based model that are important for the analysis of actuarial and insurance data in five points: Tree-based models are considered as nonparametric models and therefore do not require distributional assumptions, tree-based models can be used as a practical algorithm that can handle missing data and categorical variables in a natural way, tree-based models can automatically detect non-linear effects and potential effects, and they are easy to interpret by visualizing the tree structure in a graph, especially for smaller size trees. These advantages are particularly useful for reporting models used in actuarial and insurance data analysis.

Regression trees are employed in estimating insurance premiums by capturing the non-linear relationships between policyholder attributes and expected claim amounts. This aids insurers in setting accurate premium rates based on a nuanced understanding of risk factors. Regression trees are especially ideally suited for situations where the impact of variables is not constant because, in contrast to typical linear models, they are able to capture non-linear correlations in data. Regression trees' intuitive design makes interpretation simple, which makes it easier to communicate findings. Actuaries can better prioritize elements that have a major impact on the outcomes of interest by using regression trees, which offer insights into the relative relevance of various variables.

The combination of regression trees and machine learning has become a revolutionary force in the dynamic field of actuarial science, revolutionizing the way actuaries approach risk assessment and predictive modeling.

Machine learning augments the predictive power of regression trees, enabling the model to capture intricate relationships and dependencies within the data. This enhanced modeling capability is particularly valuable in estimating premium, predicting claim occurrences and assessing severity with a higher degree of accuracy. This study examines the mutually beneficial relationship that exists between regression trees and machine learning, highlighting the special advantages and potential uses that result from this potent union. The combination of regression trees and machine learning is set to define the forefront of data-driven decision-making in actuarial practice as the discipline continues to embrace technological breakthroughs.

Due to its increasing impact and importance in recent years, there have been many studies on classification and regression trees driven by machine learning. However, few papers can be found in the insurance literature related to regression trees and machine learning. Gardner et al. [13] use regression trees and two-stage screening were assessed by contrasting their accuracy with traditional actuarial techniques. Steadman et al. [14] proposed that a classification tree approach and two decision thresholds can enhance the use of actuarial violence risk assessment tools in clinical practice. Guelman [15] compares gradient-boosted trees with GLMs to forecast the cost of vehicle accident losses for at-fault claims. William [16] suggests a two-phase modeling process that expands on previous statistical tools like classification and regression trees, generalized linear mixed models, and actuarial methods from conventional insurance claim cost modeling. Wuthrich and Buser [17] applied various statistical methods and machine learning techniques for non-life

insurance pricing, including regression trees, bagging, random forest, boosting, and support vector machines. Diao and Weng [18] merge machine learning methods with credibility theory and suggest an approach based on regression trees to incorporate covariate data into the estimation of the credibility premium. Baillargeon [19] presents a neural architecture that can predict actuarial risk factors in accident descriptions using dense embeddings, yielding more performing and interpretable models than traditional actuarial data mining methods. Tober [20] focuses on creating and assessing three tree-based machine learning models to forecast the frequency of claims, advancing from straightforward decision trees to more complex ensemble techniques like random forests and gradient boosting machines. Henckaerts et al. [21] concentrate on using machine learning techniques to create comprehensive tariff plans based on the severity and frequency of claims. Rokicki [22] proposed the modified actuarial credibility approach, which provides accurate initial cost estimates for transport infrastructure projects, outperforming more complex methods like regression analysis and machine learning. Richman [23, 24] looks into the potential evolution and adaptation of actuarial science to include machine learning. Wong [25] provides the state of the art in ratemaking and reserving and examines how machine learning is being applied to the field of actuarial science. Quag [26] examines the various applications of tree-based models in insurance and actuarial science.

Resampling techniques play a pivotal role in machine learning, offering a strategic approach to mitigate bias, enhance model robustness, and provide a more accurate assessment of a model's performance. Among resampling processes, the "stratified random sampling" method is superior to the balanced (representative) sample when used appropriately. Stratification is the process of dividing the population into homogeneous subgroups prior to sampling.

Health insurance is more sensitive to individual characteristics, causing concentrations or infrequent conditions to be observed in the relevant risk factors and sub-fractures. For this reason, the risk factors and the probability of observation in subcategories should be taken into consideration in order to better represent the whole data in the analysis. In this study, stratified sampling was used because a non-homogeneous structure was also observed in the subgroups of various risk factors in the data examined. The theoretical background regarding stratified sampling can be found in these studies: Neyman [27], Neyman and Pearson [28], Singh and Mangat [29], and Parsons [30]. The references regarding the application of machine learning to stratified sampling are located in Liberty et al. [31], Ye et al. [32], Yu et al. [33], and Lu et al. [34].

Although actuarial science has a wealth of traditional approaches, there is a notable lack of comprehensive research on regression trees and more general machine learning applications. The absence of research in this area is especially noteworthy considering the opportunity these methods offer to improve the accuracy and flexibility of premium estimating algorithms. The new research aims to bridge this gap by delving into the unexplored area of regression trees and machine learning applications within actuarial science. Therefore, this study aims to reflect the differences between regression trees for premium forecasting from a general perspective and when they are used for forecasting purposes in the light of prior information obtained from GLM.

The remainder of the paper is organized as follows into five sections: Section 1 provides a brief introduction and a concise literature review on generalized linear models (GLM), regression trees, and the broader landscape of machine learning. Section 2 delves into the intricacies of GLM, shedding light on its mathematical foundations and highlighting its applications in actuarial science. Section 3 shifts focus to regression trees, providing a nuanced discussion on both standard regression trees (CART) and regression trees integrated with machine learning techniques. Section 4 presents a numerical analysis of premium estimation on health insurance data in the context of GLM, regression trees, and machine learning. Section 5 summarizes the results of this work and draws conclusions.

2. Generalized linear models

Generalized linear models (GLMs) are a class of statistical models that expand the linear model framework to handle a wider range of data distributions and connections. Conventional linear models make the assumption that the response variable, also known as the dependent variable, is normally distributed and that there is a linear relationship between the predictors, or independent variables, and the response. The response variable may have any of the exponential family distributions—normal, binomial, Poisson, gamma, and so on—according to GLMs, which loosen these requirements. A GLM's essential elements consist of the random component, the systematic component (linear predictor), and the link function. The response variable's probability distribution is described by the random component. It is a member of the exponential distribution family. The linear combination of the predictor variables is represented by the systemic component, also known as the linear predictor. A connection function connects it to the random component. The systematic component is connected to the expected value of the response variable by a link function. It guarantees that the model accurately captures the connection between the predictors and the distribution mean. The choice of link function depends on the distribution of the response variable. For each ith of n independently collected observations, a random component that, given the values of the explanatory variables in the model, specifies the conditional distribution of the response variable, Yi. A distribution like the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions, or the Y_i distribution, was included in the initial definition of GLMs. A linear predictor, or a linear function of regressors,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

An invertible and smooth linearizing link function, g(.) is used to convert the response variable's expectation, $\mu_i = E(Y_i)$, into a linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where g(.) is the link function; μ_i is the expected value of the response variables; $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients of the model and $x_{i1}, x_{i2}, \ldots x_{ik}$ are the predictor variables [35].

Common link functions include the identity link for Gaussian distribution, the logit link for binomial distribution, and the log link for Poisson distribution. The random component describes the response variable's probability distribution. The distribution in question is a member of the exponential family. GLMs are especially well-suited for actuarial applications where risk events frequently follow non-normal distributions, as, in contrast to linear models, they can incorporate a range of probability distributions and accommodate non-normal distributions of response variables.

GLMs are employed in this study to determine the importance and influence of the variables in our dataset. Beyond traditional linear models, GLMs offer a robust analytical framework that can handle a wide range of data distributions and capture complex interactions between variables. The variables determined to be important by GLM will be used in establishing the regression tree model, and it will be examined whether the criteria that are important in the predictions obtained according to the regression tree are similar to the variables obtained by GLM.

3. Decision Trees

A decision tree is a data analysis and machine learning tool for graphical representation and predictive modeling. It resembles an inverted tree, with each node standing for a choice or test on a certain attribute, each branch for the decision's result, and each leaf node for the outcome that was ultimately expected or the class label. Decision trees are especially helpful for tasks involving classification and regression because

they assist in decomposing intricate decision-making procedures into a number of easier to understand steps that are dependent on the data's input attributes. Because they can manage both numerical and categorical data while promoting transparency in the decision-making process, they are extensively used in a variety of industries, such as marketing, finance, and healthcare. The references on decision analysis and decision trees can be found in Magee[36], Murthy[37], Keeney [38], Tjen-Sien et. al [39] and Kotsiantis[40].

There are several types of decision trees, and their variations are often designed to address specific challenges or data characteristics. CART (Classification and Regression Trees), a decision tree that can be used for both classification and regression tasks [41]; ID3 (Iterative Dichotomiser 3), which uses entropy and information gain measures to decide the best attribute to split the dataset [42]; C4.5, which is an extension of ID3 and handles both continuous and discrete data using information gain [43] , CHAID (Chi-square Automatic Interaction Detector), which uses chi-square tests to identify significant relationships between variables, used for categorical target variables [44], Random Forest, an ensemble learning method that builds multiple decision trees, combines their predictions, helps to increase accuracy and reduce overfitting [45]. In this study, numerical analysis is performed on the CART algorithm.

3.1. CART (Classification and Regression Tree - C&RT) Algorithm

The Classification and Regression Tree Analysis (CART) algorithm was developed in 1984 by Breiman, Freidman, Olshen, and Stone [41]. CART is a straightforward but effective analytical approach that assists in identifying the most "important" (based on explanatory power) variables in a given dataset. CART algorithm divide the predictor space recursively into subsets where the distribution of y is progressively more homogeneous using a binary tree [46].

In the non-parametric regression-type CART algorithm, the data is divided into nodes based on conditional binary answers to questions containing the predictor variable y for predicting continuous dependent variables with categorical and/or continuous predictor variables.

CART can statistically show which variables, in terms of variance and explanatory power, are most significant in a model or relationship. In this sense, CART offers a complex overview of the relationships between the variables in the data and can be employed as an initial stage in building a useful model or a final representation of significant correlations. CART's visual bridge between statistical rigor and interpretation helps to make relevant and valid model creation easier [47].

The CART method creates an algorithm to predict the target values by extracting decision rules from characteristics, much like other decision tree algorithms. Both qualitative and numerical data may be included in the characteristics. Breiman et al. [41], Chipman et al. [46], Verbyla [48], Clark and Pregibon [49] are recommended readings for a comprehensive overview of the CART algorithm.

3.1.1. Regression trees

Although classification and regression in tree analysis use relatively similar statistical techniques, it's crucial to understand the differences between the two. It is desired to classify the response variable, which is often binary (0–1), in order to divide the dataset into groups. Regression trees will be used when our response variable is numeric or continuous and we want to use the data to predict the outcome. In essence, a classification tree divides the data according to homogeneity; categorizing according to similar data and filtering out the "noise" makes it more "pure"—hence the idea of a purity criterion [41]. The separations in the regression tree are performed according to the "reduction of the squares of the residuals algorithm", which means that the total variance estimated for the two resulting nodes must be minimized [41], [50]. Figure 1 is a valuable illustration of this procedure [41].

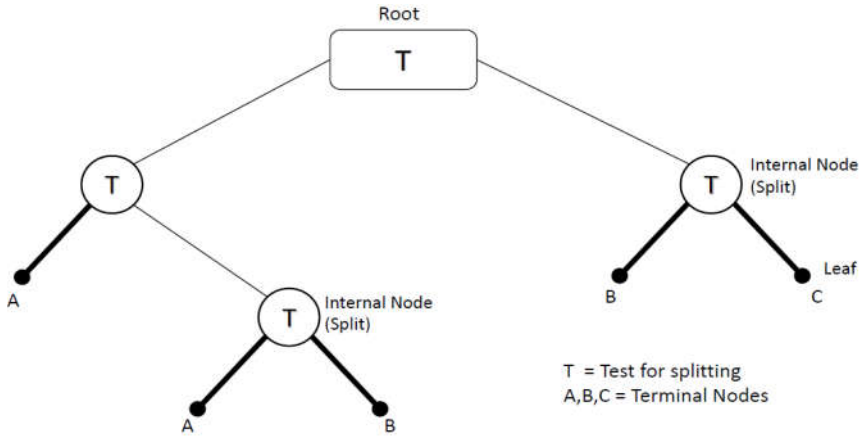

Figure 1. The structure of CART

Regression trees are an essential part of statistical modeling and machine learning, and actuarial science has greatly benefited from their capacity to reveal subtle patterns in large, complicated datasets. A tree-like model known as a regression tree divides the dataset into homogenous subsets recursively according to the values of predictor variables. Regression trees are especially useful for interpretation and prediction since each terminal node, or leaf, indicates a predicted outcome. The decision-making paths can be transparently visualized due to the tree structure's simplicity. Since CART is a non-parametric method for estimating the continuous dependent variable with categorical predictor variables and is appropriate for prediction with the variable set chosen for this investigation, it was decided to work with regression trees.

In order to improve the model's performance, adaptability, and interpretability, machine learning is integrated into regression trees through the use of sophisticated machine learning ideas and methods. In this study, it is aimed at the integration of machine learning with regression trees using splitting for training and testing data. A basic machine learning technique for assessing a model's performance is to divide the data into training and testing sets. The machine learning model (the regression tree) is trained on the training set, and its performance on untested data is assessed on the testing set. The training dataset is fed into the regression tree algorithm as part of the training process. Recursively dividing the data according to features yields decision nodes in the tree that forecast the target variable, which in regression is a continuous variable. The performance of the model must be assessed once the regression tree has been trained using the training set of data. The testing set is useful in this situation. Regression tree generalization to new, unknown data is evaluated using the testing set, which the model has not seen during training. The effectiveness of the regression tree on the testing set can be evaluated using a variety of indicators. Mean Squared Error (MSE), Mean Absolute Error (MAE), or R-squared are often used metrics in regression tasks. These metrics measure how much the actual values in the testing set depart from the projected values.

4. Application on Insurance Data

In this section, we employ a comprehensive analysis of complementary health insurance data utilizing both GLM and regression trees within a machine learning framework to evaluate the risk factors involved in the estimation of premium amounts for an insurance company.

We aim to enhance the accuracy and reliability of premium estimations, thereby catching more effective risk assessment in the estimation by using this integrated approach, harnesses the strengths of both GLM and regression trees, leveraging machine learning techniques.

Before implementation, the data set was preprocessed. Specifically, it was focused on duplicated and inconsistent data. In particular, incorrect information regarding impossible situations for employment and

the marital status of age groups was eliminated or corrected. After all corrections and data pre-processing, it was decided to conduct an examination on a sample set that would explain the entire portfolio.

All analyses were performed with the relevant packages within R programming [51].

4.1. Data

The data used in this study is complementary health insurance data from an insurance company that operates in Turkey for the period 2019-2023. The data sample was requested by the private insurance company for study purposes only. Many variables, both continuous and categorical, are included in the data sample that has been used for the analysis, as per policy. The categorical variables used in this study include the region, employment status, age group, BMI group, marital status, and gender.

Since the data set included both category and numerical variables, the features were displayed independently. Categorical features were found for every subcategory, while numerical features were represented using minimum, maximum, median, mean, and standard deviation values.

Table 1 displays the categories of categorical variables and the circumstances that were considered while assigning the classes.

Table 1. The categories of categorical variables

Category	Sub-Category	Descriptions
Region	Aegean, Black Sea, Central Anatolia, Eastern, Marmara, Mediterranean, Southeast	It is classified according to 7 main regions in Turkey
Employment status	Infant, Student, Teacher , Officer, Blue Collar, White Collar Retired, Unemployment , Other	Ages 0-6 are called infants. The majority of the ages between 7 and 20 are students.
Age group	00-06 ages; 07-20 ages; 21-25 ages; 26-30 ages; 31-35 ages; 36-40 ages; 41-45 ages; 46-50 ages; 51-55 ages; 56-60 ages; 61-65 ages; 65+	For a more specific analysis, age groups were divided into 12.
Body mass index (BMIgroup)	Infant; Underweight, Normal weight, Overweight, Obesity, High obesity	BMI groups are divided into the following categories according to the BMI range - kg/m2, World Health Organization (WHO). BMI range < 18.5 : Underweight 18.51<BMI range<24.99 : Normal weight 25<BMI range <29.99 : Overweight 30<BMI range<34.99 : Obesity BMI range > 35 : High Obesity
Marital status	Child, Single, Married, Divorced, Widow	
Gender	Female Male	

The two most important variables in premium estimation, claim amount and claim number, were included in the analysis as continuous variables. The premium amounts that were planned to be estimated and are currently used by the company were included in the analysis as dependent variables. The number and type of categorical variables on gender-based differences from the dataset are shown in the Appendix, Table A.1. The statistics of the premium amount according to the type of categorical variables based on gender are shown in Appendix A.2.

The majority of machine learning algorithms proceed on the assumption that the predictor variables are independent of each other. Mutlicollinearity, or the removal of strongly correlated predictors, is an excellent way to make an analysis robust. The correlation matrix of the continuous variables is given in Figure 2. According to Table, although there is a higher relationship between the number of claims and claim amount variables than with all other continuous variables, it is obtained that none of the variables have a significant relationship with each other.

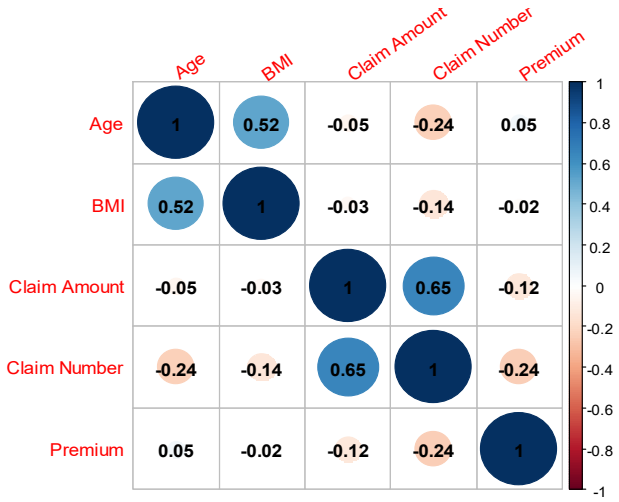

Figure 2. The correlation matrix of continous variables

4.2. Generalized Linear Model Analysis

In the GLM analysis, for the assumption of family and link functions, the premium amounts, which are the dependent variables, are visually shown to be suitable for certain distributions. In deciding which of the three available graphs is appropriate for the distribution of premium amounts, the visual consistency shown in Figure 3 is utilized. Among the default Weibull, gamma, and lognormal distributions, the gamma distribution, which is frequently used in the literature, was used together with the log link function.

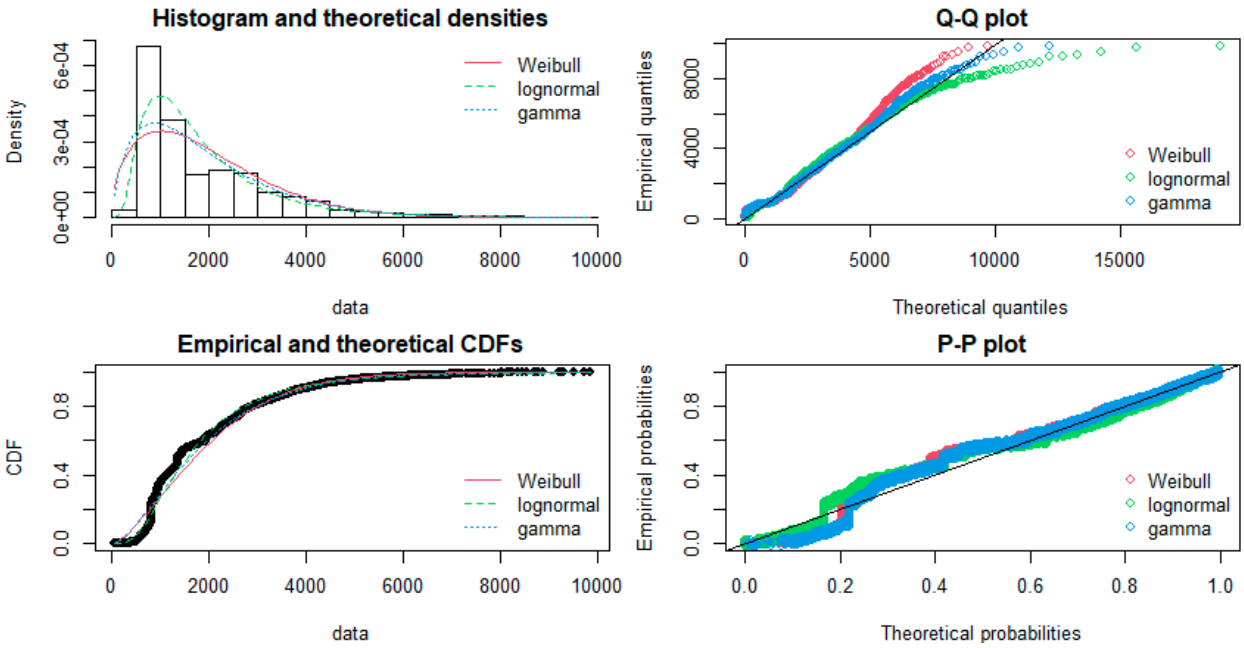

Figure 3. Histogram-density, P-P plot, Q-Q plot, and theoretical and empirical CDFs of premium amount

The GLM analysis results for the case where the premium amount is the dependent variable and all other variables are independent are shown in Table 2.

Table 2. Results of the GLM analysis

Variable	Variable level	Estimate	Std. Error	t value	p
Renewal	Renewal	0.3728	0.0253	14.7610	< 2e-16
Region	Black Sea	0.0359	0.0480	0.7490	0.4541
	Central Anatolia	-0.1025	0.0453	-2.2650	0.0235
	Eastern	-0.0643	0.0728	-0.8830	0.3770
	Marmara	0.2081	0.0409	5.0900	0.0000
	Mediterranean	0.1587	0.0597	2.6580	0.0079
	Southeast	0.0238	0.0535	0.4440	0.6571
Employment status	Infant	-2.1390	0.5798	-3.6890	0.0002
	Officer	-0.2930	0.0422	-6.9370	0.0000
	Other	-0.3204	0.0291	-11.0230	< 2e-16
	Retired	-0.0333	0.1348	-0.2470	0.8050
	Student	-0.2739	0.0711	-3.8500	0.0001
	Teacher	-0.1813	0.0764	-2.3730	0.0177
	Unemployment	0.1671	0.0907	1.8430	0.0654
	White Collar	-0.5380	0.0299	-18.0270	< 2e-16
Age	Age	-0.0100	0.0052	-1.9190	0.0550
Age Group	07-20 ages	-1.8080	0.5181	-3.4910	0.0005
	21-25 ages	-1.6720	0.5066	-3.3000	0.0010
	26-30 ages	-1.5430	0.4953	-3.1160	0.0018
	31-35 ages	-1.5200	0.4858	-3.1300	0.0018
	36-40 ages	-1.4570	0.4773	-3.0530	0.0023
	41-45 ages	-1.3930	0.4704	-2.9620	0.0031
	46-50 ages	-1.1680	0.4650	-2.5120	0.0120
	51-55 ages	-1.0230	0.4610	-2.2190	0.0265
	56-60 ages	-1.0310	0.4597	-2.2420	0.0250
	61-65 ages	-0.5294	0.5221	-1.0140	0.3106
BMI	BMI	-0.0186	0.0041	-4.4930	0.0000
BMI Group	Normal weight	-0.3237	0.1819	-1.7790	0.0753
	Obesity	0.0729	0.1938	0.3760	0.7071
	Overweight	-0.2454	0.1782	-1.3770	0.1687
	Underweight	-0.3528	0.1922	-1.8350	0.0666
Gender	Male	-0.0987	0.0192	-5.1530	0.0000
Marital Status	Divorced	0.0090	0.1116	0.0810	0.9356
	Married	0.0154	0.0957	0.1610	0.8719
	Single	-0.0171	0.0870	-0.1970	0.8440
	Widow	0.2184	0.1603	1.3630	0.1731
Claim	Claim Amount	0.0000	0.0000	5.1400	0.0000
Claim	Claim Number	-0.0756	0.0039	-19.2490	< 2e-16

The GLM, which are displayed in Table 1, indicate that the variables of employment status and region were ranked in order of significance in their respective subcategories, while the type of renewal, age group (apart from 61–65 years old), BMI, gender, claim amount, and claim number were found to be significant along with all of their subcategories. The variables age and BMI group were not found to be statistically significant, which is among the unexpected findings. The study will continue to determine whether the variables identified by GLM as significant have importance in the regression tree analysis. When the important variables are common, it will also look at how significant they are, how they are categorized into smaller groups, and how this classification affects the estimated premium amounts.

4.3.Implementation of the CART algorithm

4.3.1. General Perspective using Regression Tree

In the implementation of CART, two approaches were used in modeling the regression tree. First of all, a regression tree was created with only the variables whose importance was determined in GLM. In the second approach, modeling was applied depending on all variables in the data. Fortunately, the same results were obtained with both approaches. A visual representation of the decision tree obtained according to regression tree modeling is shown in Figure 1.

In modeling *mini split*, which refers to the minimum number of observations that are required at each node to split further, *maxdepth*, which is described as the length of the longest path from the tree root to a leaf, and the *complexity parameter (cp),* which is the minimum improvement in the model needed at each node, are applied as the control criteria. However, even possible changes in the control variables did not cause a change in the resulting regression tree.

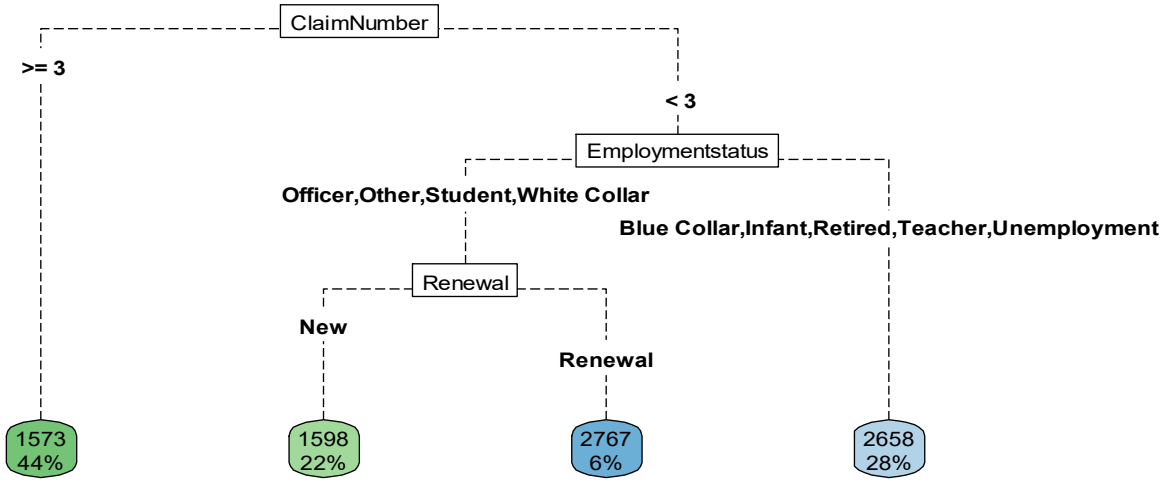

Figure 4. A visual representation of the regression tree for estimating premium amount

It can be clearly seen from Figure 4 that the premium is determined according to three important criteria in the regression tree. These variables are claim number, employment status, and renewal. These three variables are determined to be important in the GLM. However, it appears that not all variables determined to be important in GLM are taken into account in the regression tree classification.

4.3.2. Application of Regression Tree for Prediction

In the third part of the analysis, regression analysis is studied for prediction. In the analysis, first, the determination of the variable that will be the basis of the resampling method and the decision on the division ratios of the train and test data were examined.

Stratified sampling is a sampling technique used in statistical research in which the population is divided into subgroups, or "strata," based on certain characteristics, and then samples are randomly selected from each stratum. The goal is to ensure that each subgroup is represented in the sample proportionally to its presence in the overall population. Instead of separating the data into train and test as standard, we also used stratified sampling, in which the percentage of the number of observations in the categorical variables and subgroups that are important in the whole data is preserved in the selected sample. Stratified sampling is particularly useful when there are significant variations within the population and you want to ensure that each subgroup is adequately represented in the sample. This can lead to more accurate and reliable statistical

analysis, especially when dealing with diverse populations. When deciding which categorical variable to take into account in stratified sampling, the four most important categorical variables obtained from the GLM-age group, gender, employment status, and region- are taken into account. The observation and percentage densities of these four variables are also shown in Table 3.

Table 3. The observation and percentage densities of these four variables

AgeGroup		
Sub-Category	**freq**	**prob**
00-06 ages	**1023**	**20.50%**
07-20 ages	706	14.10%
21-25 ages	242	4.84%
26-30 ages	574	11.50%
31-35 ages	828	16.60%
36-40 ages	612	12.20%
41-45 ages	424	8.48%
46-50 ages	277	5.54%
51-55 ages	196	3.92%
56-60 ages	110	2.20%
61-65 ages	6	0.12%
65+	2	0.04%

Gender		
Sub-Category	**freq**	**prob**
Female	**2849**	**57%**
Male	2151	43%

Employment Status		
Sub-Category	**freq**	**prob**
Blue Collar	**1428**	**28.60%**
Infant	1023	20.50%
Officer	308	6.16%
Other	760	15.20%
Retired	24	0.48%
Student	626	12.50%
Teacher	74	1.48%
Unemployment	53	1.06%
White Collar	704	14.10%

Region		
Sub-Category	**freq**	**prob**
Aegean	276	5.52%
Black Sea	539	10.80%
Central Anatolia	854	17.10%
Eastern	110	2.20%
Marmara	**2711**	**54.20%**
Mediterranean	200	4%
Southeast	310	6.20%

Figure 5 shows the density of changes in premium amounts in relation to age groups for the variables employment status and region.

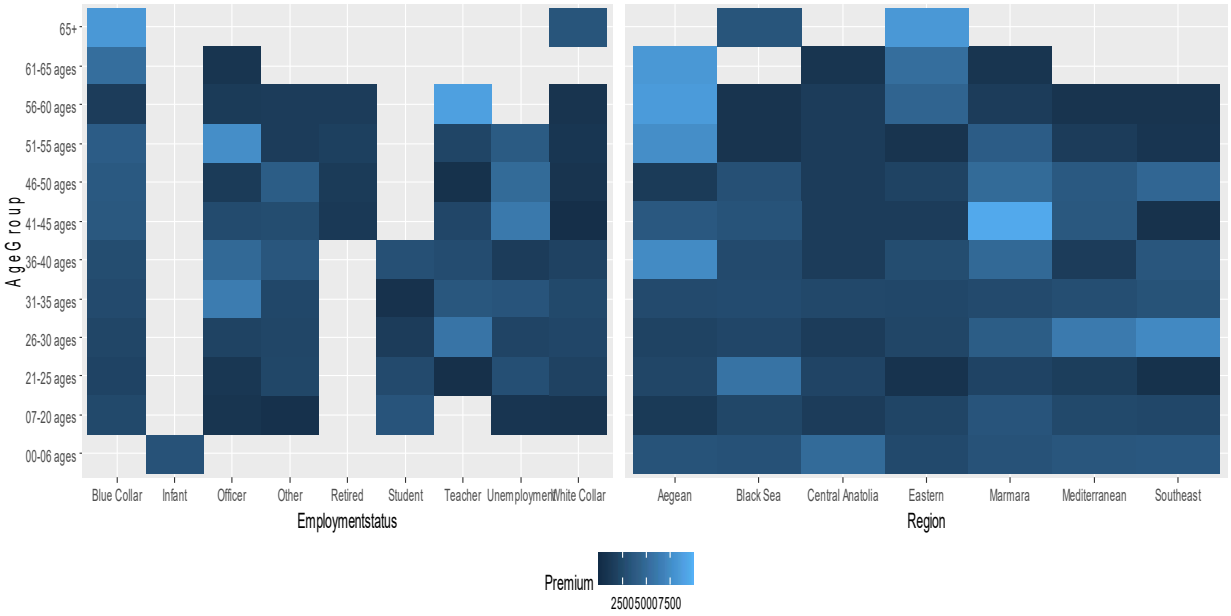

Figure 5 Changes in premium amounts in relation to age groups for the variables employment status and region.

Employment status and region have been determined to be the categorical variables with the highest degree of significance, and the effect whose percentage distribution within the sample was wanted to be explored in stratified sampling.

In regression tree algorithms, the observations should be split into training and testing data to prevent overfitting. The percentage of data used for training and testing for the validity of the model is based on the size of the dataset, the complexity of the model, and the desired performance metrics. The training dataset should be larger to have a better machine learning rate than the test dataset. Any train-test split that has more data in the training set will most likely give better accuracy as calculated on the test set. Unlike the previous standard perspective on regression trees, this section analyses train and test data in order to make predictions.

In deciding the split ratios of train and test data, the RMSE and MAE values of the prediction values were examined. One of the most successful approaches to determining the most appropriate regression tree is to decide on the appropriate split percentage by comparing the MAE and RMSEs depending on the different split ratios of the train and test data, respectively. When the performance metrics for the most commonly used split ratios of (70%–30%), (80%–20%), and (90%–10%) are compared, it is assumed that the split ratio of 70%–30%, which gives the minimum value in test errors, is appropriate and sampling is performed.

In stratified sampling, where employment status is selected as the strata, the results of the regression tree with machine learning for the estimation of premium amounts in line with the 70%–30% split ratio are displayed in Figure 6.

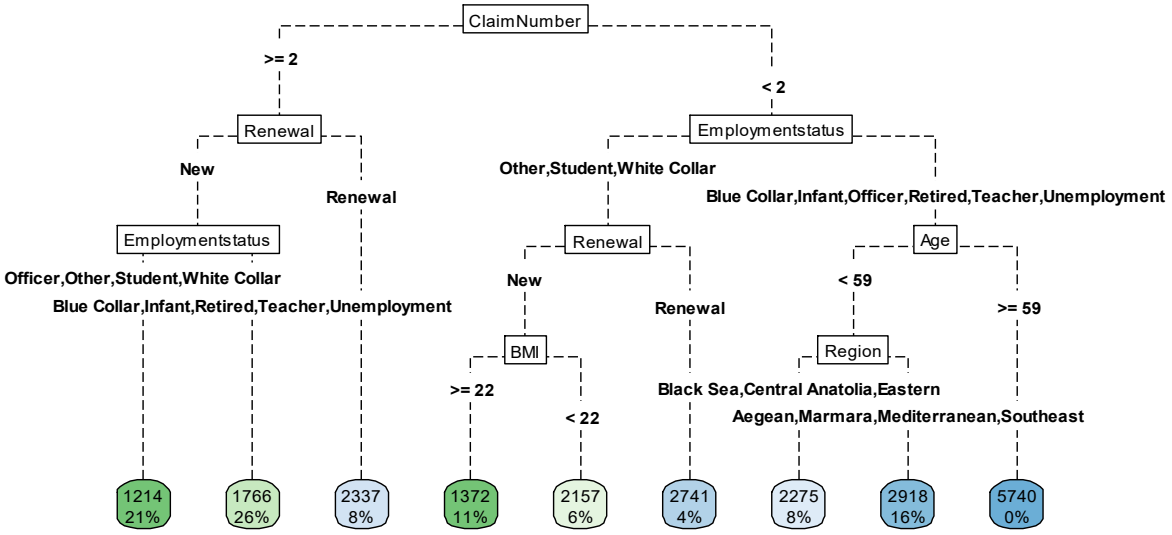

Figure 6. A visual representation of the regression tree with machine learning for estimating premium amount (Employment Status as Strata)

According to the results of the regression tree analysis, the claim number variable is the most prioritized variable, as in the general perspective using regression tree analysis. The first criterion is that the claim number is greater than or equal to 2. Then, unlike the standard perspective of regression analysis, it is determined that changes in premium amounts should be observed depending on the subcategories of renewal and employment status. Compared to the general perspective using regression tree, it is seen that BMI, age, and region variables, which are also found to be important in the GLM analysis but not included in the standard perspective regression analysis, are also important variables in the estimation of premium amounts. The results obtained show that machine learning-based regression tree analysis provides a much more comprehensive and detailed analysis by giving importance to different risk variables than general perspective using regression tree analysis.

Figure 7 presents the results of the regression tree used to estimate premium amounts in accordance with the 70%–30% split ratio in stratified sampling, where the region has been chosen as the strata.

Similar to the results in the regression tree obtained for employment status, the variability starts according to the claim number. After the claim number, employment status, renewal, age, and BMI are taken into account as variables, respectively. Surprisingly, region is not selected as a prior criterion, even in the analysis where region is selected as the base layer. The region variable was also determined as the lowest level criterion in the stratified analysis according to employment status. In fact, this analysis shows that even a variable that is determined to be important in the GLM and has high heterogeneity due to its subcategories can be evaluated after other variables in premium estimation. The result of the analysis may also vary depending on the sample size, the type of insurance, and the risk factors considered.

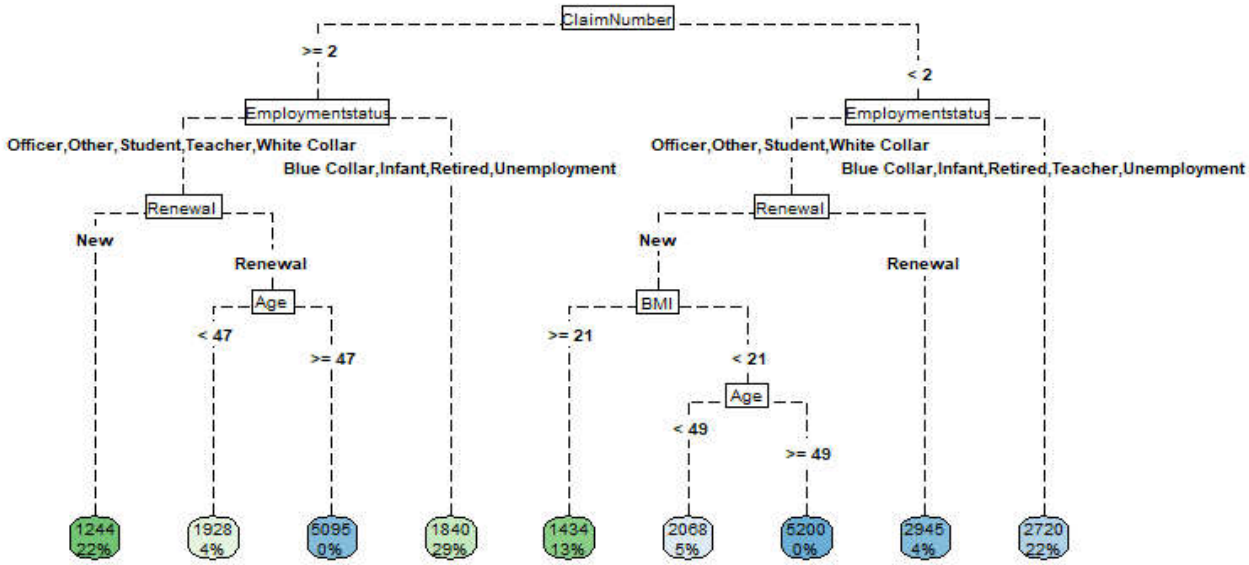

Figure 7. A visual representation of the regression tree with machine learning for estimating premium amount (Region as Strata)

A thorough investigation of the differences between the application of regression tree for prediction and general perspective using regression trees has shown that there are notable disparities in their predictive capacities. After a careful analysis, it is clear that prediction with machine learning, combining its sophisticated algorithms and methods, performs better than traditional regression trees in terms of accuracy and the capacity to pinpoint important factors.

The analysis conducted on these models indicates that machine learning-based regression trees perform more effectively by providing more accurate predictions and effectively detecting common important variables, akin to GLMs. This implies that machine learning techniques have a distinct advantage in identifying complex relationships and patterns in the data, which standard perspective regression models may struggle to discern.

One noteworthy finding is the machine learning models' proficiency in capturing variables that standard perspective regression models might overlook. The adaptability and flexibility of machine-learning algorithms enable them to handle intricate relationships and nonlinearities present in real-world datasets, leading to more nuanced and accurate predictions.

5. Conclusions and Recommendations

In the context of the non-life and health insurance sectors, the implications of adopting machine learning methodologies are particularly significant. Insurance data is often characterized by its complexity, with numerous variables interplaying to determine outcomes. Machine learning and regression trees, through their ability to handle diverse and intricate datasets, prove invaluable in accurately modeling and predicting outcomes in this sector.

This study represents a comprehensive investigation that attempts to decipher the complexity involved in risk assessment and premium calculation, starting with the fundamental GLM and ending with the creative incorporation of machine learning-driven regression trees. To begin our investigation, we first carefully looked at variables of importance using GLM, which provided a formal framework for comprehending the complex network of factors affecting health insurance premiums. The knowledge gathered from this first stage not only laid a strong foundation for further investigations, but it also highlighted how crucial variable priority is to the actuarial decision-making process. The variable importance comparison between regression trees and GLM supplied insightful information on the different viewpoints that each methodology presented, paving the way for a more in-depth comprehension of the variables influencing premium estimates. We took a step farther in terms of technological innovation and included machine learning into our regression tree method. This combination improved our model's ability to adjust to the changing and dynamic health insurance market while also improving the precision of our premium estimation.

Furthermore, this study underscores the potential for further improvement by expanding the scope of data and considering different types of insurance. The application of machine learning perspectives can be extended to other insurance sectors, fostering a more comprehensive understanding of the intricacies involved. By incorporating diverse datasets and varying insurance contexts, researchers can refine their models to enhance predictive accuracy and relevance across a broader spectrum. Regression trees and machine learning are expected to play an increasingly important role in actuarial science as technology advances, providing fresh perspectives and creative approaches to the problems associated with risk assessment and management.

Kaynaklar

[1] P. McCullagh, J. A. Nelder,1989, *Generalized Linear Models 2nd ed.*. London: Chapman and Hall.

[2] A. E. Renshaw ,1991, Actuarial graduation practice and generalized linear and non-linear models. J Inst. Act., 118, 295-312.

[3] A. E. Renshaw, P. Verrall, 1994, A Stochastic Model Underlying The Chain Ladder Technique. In *Proceedings of the XXV ASTIN Colloquium, Cannes*.

[4] S. Haberman, A. E. Renshaw, 1996, *Generalized Linear Models and Actuarial Science*. *Journal of the Royal Statistical Society. Series D The Statistician*, 454, 407–436. https://doi.org/10.2307/2988543

[5] A. J. Dobson, 2002, *An Introduction to Generalized Linear Models Second Edition*. London: Chapman and Hall/CRC.

[6] D. Andersen, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi, 2005, *A Practitioner's Guide to Generalized Linear Models Second Edition*. CAS Study Note.

[7] K. Antonio, J. Beirlant, 2007, Actuarial statistics with generalized linear mixed models. *Insurance Mathematics & Economics*, 40, pp. 58-76. https://doi.org/10.1016/J.INSMATHECO.2006.02.013.

[8] P. De Jong, G. Heller, 2008, *Generalized Linear Models for Insurance Data International Series on Actuarial Science*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511755408

[9] M. V. Wüthrich, M. Merz, 2008, *Stochastic claims reserving methods in insurance*. John Wiley & Sons.

[10] E. Ohlsson, B. Johansson, 2010, *Non-Life Insurance Pricing with Generalized Linear Models*. Springer.

[11] E. W. Frees, 2015, *Analytics of insurance markets*. *Annual Review of Financial Economics*, 7, 253–77

[12] Z. Quan, Insurance Analytics with Tree-Based Models. PhD thesis, University of Connecticut, 2019.

[13] W. Gardner, C. Lidz, E. Mulvey, E. C. Shaw, 1996, A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. *Law and Human Behavior*, 20, 35-48.

[14] H. Steadman, E. Silver, J. Monahan, P. Appelbaum, P. Robbins, E. Mulvey, T. Grisso, L. Roth, S. Banks, 2000, A Classification Tree Approach to the Development of Actuarial Violence Risk Assessment Tools. *Law and Human Behavior*, 24, 83-100. https://doi.org/10.1023/A:1005478820425.

[15] L. Guelman, 2012, Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 393, 3659–67.

[16] J. William, M. Martin, C. Chojenta, D. Loxton, 2018, *An actuarial investigation into maternal hospital cost risk factors for public patients*. *Annals of Actuarial Science*, 12, 106 - 129. https://doi.org/10.1017/S174849951700015X.

[17] M. V. Wuthrich, C. Buser, 2023, *Data Analytics for Non-Life Insurance Pricing*. Swiss Finance Institute Research Paper No. 16-68. Available at SSRN: https://ssrn.com/abstract=2870308 or http://dx.doi.org/10.2139/ssrn.2870308

[18] L. Diao, C. Weng, 2019, *Regression Tree Credibility Model*. *North American Actuarial Journal*, 232, 169-196. DOI: 10.1080/10920277.2018.1554497

[19] J. Baillargeon, L. Lamontagne, É. Marceau, 2020, *Mining Actuarial Risk Predictors in Accident Descriptions Using Recurrent Neural Networks*. *Risks*. https://doi.org/10.3390/risks9010007.

[20] S. Tober, 2020, *Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling Dissertation*. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-276233

[21] R. Henckaerts, M.-P. Côté, K. Antonio, R. Verbelen, 2021, *Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods*. *North American Actuarial Journal*, 252, 255-285. DOI: 10.1080/10920277.2020.1745656

[22] B. Rokicki, K. Ostaszewski, 2022, *Actuarial Credibility Approach in Adjusting Initial Cost Estimates of Transport Infrastructure Projects*. *Sustainability*. https://doi.org/10.3390/su142013371.

[23] R. Richman, 2021a, *AI in actuarial science—a review of recent advances—part 1*. *Ann. Actuar. Sci.*, 152, 207–29

[24] R. Richman, 2021b, *AI in actuarial science—a review of recent advances—part 2*. *Ann. Actuar. Sci.*, 152, 230–58

[25] B. Wong, J. Christopher, H. Cossette, L. Lamontagne, E. Marceau, 2021, *Machine Learning in P&C Insurance: A Review for Pricing and Reserving*. *Risks*, 91, 4. https://doi.org/10.3390/risks9010004

[26] Z. Quan, 2019, *Insurance Analytics with Tree-Based Models Doctoral Dissertations No. 2374*. Retrieved from https://digitalcommons.lib.uconn.edu/dissertations/2374

[27] J. Neyman, 1934, On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625

[28] J. Neyman, E. S. Pearson, 1933, On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337. http://www.jstor.org/stable/91247

[29] R. Singh, N. S. Mangat, 1996, *Stratified Sampling*. In: *Elements of Survey Sampling*, Vol. 15. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1404-4_5

[30] V. L. Parsons, 2014, *Stratified sampling*. *Wiley StatsRef: Statistics Reference Online*, 1-11.

[31] E. Liberty, K. Lang, K. Shmakov, 2016, June. *Stratified sampling meets machine learning*. In *International conference on machine learning* pp. 2320-2329. PMLR.

[32] Y. Ye, Q. Wu, J. Z. Huang, M. K. Ng, X. Li, 2013, *Stratified sampling for feature subspace selection in random forests for high dimensional data*. *Pattern Recognition*, 463, 769-787.

[33] T. Yu, X. Zhai, S. Sra, 2019, *Near Optimal Stratified Sampling*. *ArXiv, abs/1906.11289*.

[34] Y. Lu, Y. Park, L. Chen, Y. Wang, C. De Sa, D. Foster, 2021, July. *Variance reduced training with stratified sampling for forecasting models*. In *International Conference on Machine Learning* pp. 7145-7155. PMLR.

[35] J. Fox, 2008, *Applied Regression Analysis and Generalized Linear Models*, 2nd Edn. Thousand Oaks, CA: Sage.

[36] J.F. Magee, 1964, *Decision trees for decision making, Harvard Business Review, pp. 126-138.*

[37] S.K. Murthy, 1998, Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining Knowl Discovery* 2(4):345–389

[38] R.L. Keeney, 1982, Decision Analysis: An Overview. *Operations Research*, *30*(5).
[39] L.Tjen-Sien, L. Wei-Yin, S.Yu-Shan, 2000, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach Learn* 40:203–228
[40] S.B. Kotsiantis, 2013, Decision trees: a recent overview. *Artif Intell Rev* 39, 261–283
[41] L. Breiman, J. Friedman, R. Olshen, C. J. Stone, 1984, *Classification and regression Trees*. Wadsworth, Belmont, CA.
[42] J.R. Quinlan, 1986, Induction of decision trees. *Mach Learn 1*, 81–106.
[43] J.R. Quinlan, 1993, *C4.5: programs for machine learning*. Morgan Kaufmann, San Francisco
[44] G.V. Kass, 1980, "An Exploratory Technique for Investigating Large Quantities of Categorical Data". *Applied Statistics.* 29 (2): 119–127
[45] J. Gehrke, R. Ramakrishnan, V. Ganti, 2000, RainForest: a framework for fast decision tree construction of large datasets. *Data Mining Knowl Discovery* 4(2–3):127–162
[46] H. A. Chipman, E. I. George, R. E. McCulloch, 1998, *Bayesian CART model search*. *Journal of the American Statistical Association*, 93443, 935-960 pp.
[47] J. Morgan, 2014, *Classification and regression tree analysis*. Boston: Boston University, 298.
[48] D. L. Verbyla, 1987, *Classification trees: a new discrimination tool*. *Canadian Journal of Forest Research,* 17, 9, 1150–1152.
[49] L. A. Clark, D. Pregibon, 1992, *Tree-based models*. In: *Statistical models* Eds. Chambers JM, Hastie TJ. Pacific Grove, CA: Wadsworth, p 377–419.
[50] G. De'ath, K. E. Fabricius, 2000, *Classification and Regression Trees: A Powerful yet Simple* Technique for Ecological Data Analysis. Ecology, 81, 3178-3192
[51] R Core Team , 2021, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. RStudio 2023.09.1

Appendix

Table A.1 The number and the type of categorical variables on gender-based

Categories	Sub-categories	Gender				
		Female		Male		Total
		Number of Observations	Percentage (%)	Number of Observations	Percentage (%)	Number of Observations
Renewal	New	2423	57,4%	1796	42,6%	4219
	Renewal	426	54,5%	355	45,5%	781
Region	Aegean	154	55,8%	122	44,2%	276
	Black Sea	317	58,8%	222	41,2%	539
	Central Anatolia	506	59,3%	348	40,7%	854
	Eastern	49	44,5%	61	55,5%	110
	Marmara	1560	57,5%	1151	42,5%	2711
	Mediterranean	113	56,5%	87	43,5%	200
	Southeast	150	48,4%	160	51,6%	310
Employment statue	Blue Collar	947	66,3%	481	33,7%	1428
	Infant	489	47,8%	534	52,2%	1023
	Officer	188	61,0%	120	39,0%	308
	Other	445	58,6%	315	41,4%	760
	Retired	10	41,7%	14	58,3%	24
	Student	311	49,7%	315	50,3%	626
	Teacher	56	75,7%	18	24,3%	74
	Unemployment	33	62,3%	20	37,7%	53
	White Collar	370	52,6%	334	47,4%	704
Age group	00-06 ages	489	47,8%	534	52,2%	1023
	07-20 ages	347	49,2%	359	50,8%	706
	21-25 ages	163	67,4%	79	32,6%	242
	26-30 ages	403	70,2%	171	29,8%	574
	31-35 ages	540	65,2%	288	34,8%	828
	36-40 ages	335	54,7%	277	45,3%	612
	41-45 ages	250	59,0%	174	41,0%	424
	46-50 ages	154	55,6%	123	44,4%	277
	51-55 ages	110	56,1%	86	43,9%	196
	56-60 ages	54	49,1%	56	50,9%	110
	61-65 ages	3	50,0%	3	50,0%	6
	65+	1	50,0%	1	50,0%	2
BMI group	High obesity	7	50,0%	7	50,0%	14
	Infant	489	47,8%	534	52,2%	1023
	Normal weight	1827	64,1%	1022	35,9%	2849
	Obesity	15	29,4%	36	70,6%	51
	Overweight	329	43,8%	422	56,2%	751
	Underweight	182	58,3%	130	41,7%	312
Marital status	Child	783	48,0%	849	52,0%	1632
	Divorced	96	73,3%	35	26,7%	131
	Married	1639	62,0%	1003	38,0%	2642
	Single	307	54,0%	262	46,0%	569
	Widow	24	92,3%	2	7,7%	26
Gender	Male	0	0,0%	2151	100,0%	2151
	Female	2849	100,0%	0	0,0%	2849

Table A.2 The statistics of the premium amount according to type of categorical variables on gender-based

		Gender							
		Female				Male			
		Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation
Region	Aegean	408,87	8706,76	1990,77	1817,32	421,00	8383,63	1712,23	1291,11
	Black Sea	207,04	9229,54	1899,28	1471,04	381,94	5279,54	1591,23	1043,38
	Central Anatolia	58,92	8451,76	1580,52	1268,81	170,65	9742,52	1537,22	1154,88
	Eastern	509,29	6255,22	1775,35	1145,54	502,81	8171,71	1857,61	1416,85
	Marmara	128,45	9846,17	2163,27	1544,13	121,44	8720,90	2051,50	1395,82
	Mediterranean	279,91	7656,97	2392,99	1857,63	400,00	6753,05	1849,52	1334,71
	Southeast	400,00	7170,24	2092,06	1561,73	422,81	9301,10	1607,51	1334,53
Employment status	Blue Collar	162,34	9846,17	2495,51	1776,09	381,94	8171,71	2136,76	1392,46
	Infant	80,60	8021,25	2016,56	1353,37	157,55	9301,10	2052,79	1405,56
	Officer	197,67	8706,76	1726,85	1593,68	170,65	8383,63	1399,38	1004,46
	Other	66,33	8803,85	1686,88	1356,41	148,09	8187,70	1707,33	1291,06
	Retired	1211,75	5072,30	2764,98	1462,63	800,00	9742,52	2756,14	2489,18
	Student	58,92	4921,72	1706,66	1005,75	121,44	5597,49	1767,53	1075,69
	Teacher	306,99	7971,17	1872,25	1570,70	732,76	8720,90	2481,27	1873,05
	Unemployment	549,32	6121,14	2355,29	1516,25	1123,09	6017,00	2903,40	1281,82
	White Collar	532,97	9508,34	1590,98	1259,57	400,00	8041,72	1383,25	1149,38
Age group	00-06 ages	80,60	8021,25	2016,56	1353,37	157,55	9301,10	2052,79	1405,56
	07-20 ages	58,92	4921,72	1664,83	993,24	121,44	5597,49	1770,35	1074,25
	21-25 ages	162,34	9229,54	2085,46	1634,46	400,00	4187,46	1434,33	869,77
	26-30 ages	197,67	8538,31	2084,77	1766,13	170,65	4621,44	1467,33	980,19
	31-35 ages	66,33	8706,76	1917,77	1433,89	148,09	5614,11	1597,07	1030,86
	36-40 ages	155,91	9508,34	2016,92	1512,32	563,75	5801,39	1671,15	1153,47
	41-45 ages	595,00	9290,45	2147,73	1530,49	155,43	8187,70	1837,66	1400,53
	46-50 ages	207,04	8803,85	2239,93	1709,49	595,00	6790,11	2286,24	1579,41
	51-55 ages	400,00	8536,70	2505,87	2114,11	595,00	8041,72	2405,58	1822,79
	56-60 ages	800,00	9846,17	2377,62	2110,55	344,20	9742,52	2495,62	2287,57
	61-65 ages	850,00	8164,46	3288,15	4223,01	4382,26	5586,93	5065,45	618,40
	65+	3372,89	3372,89	3372,89	.	8171,71	8171,71	8171,71	.
BMII group	High obesity	408,87	4313,85	1923,88	1559,13	1315,24	5381,07	2408,08	1455,94
	Infant	80,60	8021,25	2016,56	1353,37	157,55	9301,10	2052,79	1405,56
	Normal weight	128,45	9508,34	1987,61	1546,90	121,44	9742,52	1745,83	1278,02
	Obesity	472,77	7663,38	2905,90	2227,08	589,22	6790,11	2374,88	1635,28
	Overweight	66,33	9846,17	2209,16	1755,73	155,43	8720,90	1854,93	1364,45
	Underweight	58,92	9290,45	1939,44	1289,66	421,00	5279,54	1724,93	1085,18
Marital status	Child	58,92	8021,25	1897,84	1247,26	121,44	9301,10	1943,42	1291,46
	Divorced	595,00	8472,84	2249,49	1682,69	525,68	5283,10	1892,55	1261,13
	Married	66,33	9508,34	2043,88	1631,43	148,09	9742,52	1833,54	1420,07
	Single	480,32	8803,85	2032,34	1414,23	400,00	5802,25	1630,47	1042,95
	Widow	595,00	9846,17	3270,39	2748,35	1478,53	5897,23	3687,88	3124,49
Gender	Female	58,92	9846,17	2019,76	1532,21	.	.	.	.
	Male	.	.	.	.	121,44	9742,52	1854,86	1331,16