DEGREE PROJECT IN MATHEMATICS,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2017

Risk Premium Prediction of Car Damage Insurance using Artificial Neural Networks and Generalized Linear Models

LOVISA STYRUD

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Risk Premium Prediction of Car Damage Insurance using Artificial Neural Networks and Generalized Linear Models

LOVISA STYRUD

Degree Projects in Mathematical Statistics (30 ECTS credits)
Degree Programme in Applied and Computational Mathematics (120 credits)
KTH Royal Institute of Technology year 2017
Supervisors at If Skadeförsäkring: Jonna Alnervik and Bengt Eriksson
Supervisor at KTH: Jimmy Olsson
Examiner at KTH: Jimmy Olsson

TRITA-MAT-E 2017:28
ISRN-KTH/MAT/E--17/28--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

Over the last few years the interest in statistical learning methods, in particular artificial neural networks, has reawakened due to increasing computing capacity, available data and a strive towards automatization of different tasks. Artificial neural networks have numerous applications, why they appear in various contexts. Using artificial neural networks in insurance rate making is an area in which a few pioneering studies have been conducted, with promising results. This thesis suggests using a multilayer perceptron neural network for pricing car damage insurance. The MLP is compared with two traditionally used methods within the framework of generalized linear models. The MLP was selected by cross-validation of a set of candidate models. For the comparison models, a log-link GLM with Tweedie's compound Poisson distribution modeling the risk premium as dependent variable was set up, as well as a two-parted GLM with a log-link Poisson GLM for claim frequency and a log-link Gamma GLM for claim severity. Predictions on an independent test set showed that the Tweedie GLM had the lowest prediction error, followed by the MLP model and last the Poisson-Gamma GLM. Analysis of risk ratios for the different explanatory variables showed that the Tweedie GLM was also the least discriminatory model, followed by the Poisson-Gamma GLM and the MLP. The MLP had the highest bootstrap estimate of variance in prediction error on the test set. Overall however, the MLP model performed roughly in line with the GLM models and given the basic model configurations cross-validated and the restricted computing power, the MLP results should be seen as successful for the use of artificial neural networks in car damage insurance rate making. Nevertheless, practical aspects argue in favor of using GLM.

This thesis is written at If P&C Insurance, a property and casualty insurance company active in Scandinavia, Finland and the Baltic countries. The headquarters are situated in Bergshamra, Stockholm.

Sammanfattning

De senaste åren har det skett en dramatisk ökning av intresset för metoder inom statistisk inlärning, speciellt artificiella neurala nät. Anledningar till detta är ökad datorkapacitet och tillgänglig data samt en önskan om att effektivisera olika typer av uppgifter. Artificiella neurala nät har en mängd olika tillämpningsområden och återfinns därför i olika kontexter. Användandet av artificiella neurala nät för prissättning av försäkringar är ett område inom vilket det har utförts ett antal inledande studier med lovande resultat. I den här masteruppsatsen används en multilayer perceptron för att prissätta vagnskadeförsäkring och jämförs med två vanliga metoder för prissättning genom generaliserade linjära modeller. MLP-modellen valdes ut genom korsvalidering av en uppsättning tänkbara modeller. För jämförelse sattes en GLM-modell med logaritmisk länkfunktion och Tweedies sammansatta poissonfördelning upp där den beroende variabeln utgörs av riskpremien, samt en tvådelad GLM-modell innefattande en poissonfördelad GLM med logaritmisk länk för skadefrekvensen och en gammafördelad GLM med logaritmisk länk för skadestorleken. Prediktioner på oberoende testdata visade att Tweedie GLM-modellen hade det lägsta prediktionsfelet följt av MLP-modellen och sist Poisson-Gamma GLM-modellen. Analys av riskkvoter för de olika förklarande variablerna visade att Tweedie GLM-modellen också var den minst diskriminerande modellen, följt av Poisson-Gamma GLM-modellen och MLP-modellen. MLP-modellen hade den högsta bootstrappade uppskattningen av prediktionsfelet på testdatat. På det hela taget visade dock MLP-modellen resultat ungefär i linje med GLM-modellerna och givet de enkla nätverksstrukturer som korsvaliderats samt begränsningen i datorkapacitet bör ändå MLP-resultaten ses som en framgång för användandet av neurala nät inom prissättning av vagnskadeförsäkring. Dock finns det stora praktiska fördelar med generaliserade linjära modeller.

Denna masteruppsats har skrivits för If Skadeförsäkring, ett försäkringsbolag med kunder i Skandinavien, Finland och Baltikum. Huvudkontoret ligger i Bergshamra, Stockholm.

Acknowledgements

I want to start by thanking Jimmy Olsson, my supervisor at KTH, for continuous guidance and advice. I also want to thank Bengt Eriksson, Jonna Alnervik, Hanna Nyquist and Vilhelm Luttemo at If for the idea behind the thesis and valuable counseling. Finally, I want to thank David Ödling for always being there. You are the best.

Insurance Terminology

gross premium	the price of an insurance contract
risk premium	part of premium corresponding to the insurance risk
policyholder	buyer of an insurance contract
insurer	issuer of an insurance contract, often an insurance company
insurance risk	probability that the insurer is obliged to pay the policyholder due to occurrence of insured events, defined by the insurance contract between the insurer and the policyholder
claims cost	sum of payments to the policyholder from the insurer due to occurrence of insured events
exposure	period of time during which an insurer is exposed to insurance risk
rate making	actuarial work of determining adequate premiums
claim frequency	number of insurance claims per time period
claim severity	cost per incurred claim

Table of Contents

1 Introduction

The core business of insurance companies is to sell contracts protecting the insureds from economic stress in the case of unexpected events. The amount and circumstances under which an insured is to receive economic compensation is defined in the agreement between the insurer and the insured. Common types of insurance for private individuals are home and auto insurance.

A key issue is how to price an insurance contract, which is also known as rate making. If the price is too high, customers will turn to other insurance companies, and if the price is too low, the insurance company will not receive enough premiums to cover the insureds' claim costs. Also, it seems reasonable to charge different premiums to different customers based on some well-chosen variables which are correlated to the insurance risk of the specific customer. In auto insurance, it could be that the risk is correlated to the brand of the car the insureds drive or to how many years the insureds have had their driving licenses. The part of the gross premium corresponding to the insurance risk is known as the risk premium, which is hence the expected claim cost. On average, an insurance company needs to charge more than the risk premium, since costs for administration need to be covered and a profit is often wanted. Note that the gross premium charged could also depend on price optimization strategies based on e.g. price elasticity.

Obviously, understanding the insurance risk of each contract is absolutely essential in an insurance business. If the risk is not understood, the profitability of the insurance company could decrease or the company might not even be able to meet its liabilities.

Traditionally linear regression models have been used to model the risk premium. During the past decades, there has been a transition towards using generalized linear models, GLM, since these types of models have shown to be more suitable for rate making than linear regression models. However, the use of GLM has some potential drawbacks. First, the distribution of the output needs to be specified. Also, GLM are not suited for modeling high-dimensional nonlinear dependencies between explanatory variables since interaction effects between explanatory variables need to be manually included in the model.

Recently, there has been a reawakened interest and development of methods in statistical learning, particularly in artificial neural networks. These can

be designed to have the desired ability of modeling sophisticated nonlinear dependencies in data. The question then arises whether a well-chosen artificial neural network could be able to predict expected claim costs more accurately than GLM.

1.1 Previous work

A number of studies on the use of neural networks for prediction of risk premiums have been conducted, with promising results. In one of the larger, several statistical learning methods for pricing car insurance were compared (Dugas et al. 2003). Models from the families of linear regression, generalized linear models, mixture models, decision trees, artificial neural networks and support vector machines were fitted to car insurance data with the purpose of predicting claim amounts given a certain set of input variables. The same 33 explanatory variables were used when fitting all models. The claims in the data were from bodily injury, accident benefit, property damage, collision, comprehensive (i.e. theft, vandalism, fire etc.), death benefit and loss of use. The models were compared with an intercept model as benchmark, which is the mean of all claim amounts. The results show that both of the two neural network models tested had a lower mean squared error, MSE, on both the validation and test sets than the GLM. The lowest validation and test MSE were seen for a mixture model. The training, validation and test MSE for all the models were rather similar, which, according to the authors, is due to the heavy right tail of the claim distribution. However, the authors conclude by arguing for the use of neural networks to estimate risk premiums in car insurance.

Mano and Rasa compare GLM, neural networks and decision trees for modeling risk premiums in personal insurance (Mano and Rasa 2012). The authors highlight that GLM is well suited for insurance rate making because of how well such models can be tuned to suit insurance data and their ability to handle large amounts of data. The authors also stress the problem that neural networks can take vast amounts of time to train on a data set, even though they are confident that a neural network can be as good as a GLM model for predicting risk premiums. Furthermore, they think that with neural networks or decision trees, it is not necessary to model claim frequency and claim severity separately, as is often done with GLM.

In 'Neural Networks Demystified', Francis stresses that artificial neural networks are universal function approximators as well as a tool for variable

selection (Francis 2001). She demonstrates how to fit a multilayer perceptron to realistically simulated car insurance data stretching over 6 years, with the aim of predicting claim severity. The claim amounts were drawn from a lognormal distribution, using a scale parameter μ which is dependent on the characteristics of the policyholder. Among the explanatory variables were driver age, car type (4 groups), car age, territory (45 groups), credit (represents creditworthiness of policyholder) and a number of inflation factors. Two of the explanatory variables contained missing data; car age and credit. The data set comprised 5000 observations, each observation representing an individual policyholder. A training set was created from 4000 of the observations. The remaining 1000 observations were put aside for testing. MLP models with one hidden layer comprising $3, 4, 5, 6$ and 7 nodes were fitted to the training data set, with the log of claim severity as dependent variable. The model with 4 hidden nodes performed the best on the test set, in the sense that it had the highest $R^2 = 5\%$. Francis admits that it is a low value, but argues that it is due to a high degree of randomness in the data. The model showed good ability of identifying high and low claim severities. The neural network model was compared to a linear regression model with explanatory variables chosen with forward stepwise selection. The comparison showed that the regression model had a lower R^2 than the neural network model, although Francis says that for some measures of goodness of fit, the regression model had almost as good results as the neural network model.

In another study, a five-layer fuzzy adaptive neural network was constructed to model the total claim amount using data from a Turkish insurance company (Dalkilic et al. 2009). *Fuzzy* refers to fuzzy set theory, in which observations have degrees of membership in different sets, ranging from 0 to 1 (Zadeh 1965). A fuzzy clustering algorithm was applied to the observations before training the network. The sum of squared errors, SSE, for the prediction of the total claim amount was $\text{SSE}_{NN} = 0.0207$ for the neural network, compared to $\text{SSE}_{LS} = 2.2392$ for an ordinary multiple linear regression model. Note that it is unclear how training and validation has been performed. Explanatory variables were the total number of claims and ordinal number of the calendar month.

1.2 Objectives

The results from previous studies of using artificial neural networks in insurance rate making are encouraging, and suggest more and deeper analysis of how neural networks compare to more traditional modeling techniques with

GLM. This application of neural networks is still fairly new, and further studies in different insurance fields are required to investigate the potential of neural networks in a real-world insurance business.

The objective of this thesis is to study how a multilayer perceptron, which is a type of artificial neural network, compares to GLM for modeling the risk premium in car damage insurance. The idea is that a well-chosen multilayer perceptron, MLP, has the ability to model high-dimensional nonlinearities between explanatory variables and should thus produce less biased fits on training data compared to the less flexible GLM. Thus, provided a suitable model configuration and enough time available for training, this thesis will investigate whether an MLP could have a lower error on an independent test set than a GLM for this specific insurance problem. It will also be investigated how an MLP compares to GLM in the sense of fairness in charged risk premiums between different groups of policyholders. Since knowing the policyholders' risks is of utmost importance in insurance, understanding of how artificial neural networks compare to GLM in rate making is attractive for the industry. The purpose of the thesis is also to analyze how well-suited a neural network model is for the problem from a perspective of practical implementation.

1.3 Disposition

In Section 2, a mathematical background of generalized linear models and artificial neural networks is presented, as well as the model assessment methods mean squared error, cross-validation, bootstrap, risk ratio evaluation and sensitivity analysis. In Section 3, the method of the study is presented. The section starts with a presentation of the data preprocessing which includes choice of explanatory variables and analysis of the claim size distribution. The two models within the GLM framework chosen for comparison with the neural network model are then presented: the Tweedie GLM with risk premium as dependent variable and the Poisson-Gamma GLM where claim frequency and severity are modeled separately. Then the neural network modeling process is described. This includes choice of activation functions and choice of network architecture. In Section 4, the results of the study are presented. The neural network cross-validation results are given in Section 4.1. These are the results upon which the choice of final neural network model is based. In Section 4.2, the cross-validation results for the Tweedie GLM are presented. Section 4.3 comprises the results for the comparison between the neural network model, the Tweedie GLM and the Poisson-Gamma GLM.

The models are compared by means of mean squared error on an independent test set, risk ratios for subsets of policyholders, risk ratios for different magnitudes of claim size, bootstrap estimates of variance for the different models and sensitivity of the explanatory variables. In Section 5, the results are discussed from the perspective of the objectives for the study. Also, the results are discussed from the perspective of practical implementation and use. Moreover, method improvements and future work are discussed. The thesis is concluded in Section 6.

2 Mathematical background

2.1 Risk premium

Let $X \in \mathcal{R}^p$ be a stochastic $1 \times p$ vector with a policyholder specific set of explanatory variables such as age, population density at place of residence, car brand etc. The number of explanatory variables is hence p. Let $A \in \mathcal{R}^+$ be a stochastic variable for the corresponding claim cost.

The risk premium is then defined as (Dugas et al. 2003)

$$f(x) = E[A|X = x] \ . \tag{1}$$

2.2 Generalized linear models

Generalized linear models, GLM, is a class of models which is a generalization of classical linear models. Recall that a classical linear model is on the form

$$y = \bar{X}\beta + e \ , \tag{2}$$

where y is an $n \times 1$ vector comprising n observations y_i, $i = 1, \ldots, n$, of the dependent variable, $\bar{X}$ is an $n \times p$ matrix comprising observations of the p explanatory variables, β is an $p \times 1$ vector of coefficients and e is an $n \times 1$ vector of error terms.

In a classical linear model, the observations of y are seen as outcomes of a random $n \times 1$ vector Y. The elements of Y are assumed to be independent and normally distributed and are also assumed to have constant variance, i.e. $Y_i \in N(\mu_i, \sigma^2)$. Furthermore, we have that $E(Y) = \bar{X}\beta = \mu$, where $\bar{X}\beta$ is the systematic part of (2). Moreover, $\eta = \bar{X}\beta$ is called the linear predictor of (2) and we have that $\mu = \eta$ (McCullagh and Nelder 1983).

In generalized linear models, the distribution of the elements of Y is allowed to be any distribution within the exponential family of distributions (McCullagh and Nelder 1983). These include e.g. the binomial, Poisson, normal and gamma distributions, and are on the form

$$h(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{3}$$

for some functions a, b, c, and θ, which is known as the canonical parameter and is related to the mean of the specific distribution (Olsson 2002).

Furthermore, we let $g(\mu) = \eta = \bar{X}\beta$, where $g(\cdot)$ is known as the link function which is supposed be monotone and differentiable. Hence,

$$E[y] = \mu = g^{-1}(\eta) = g^{-1}(\bar{X}\beta) \ . \tag{4}$$

For the classical linear model, g is simply the identity link and $\eta = \mu$. In insurance, the exponential link is often used (Dugas et al. 2003), giving a model $\hat{f}(x)$ on the form

$$\hat{f}(x) = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_i\right) \tag{5}$$

for the risk premium for a policyholder with $X = x$.

This model has the tractable property that $\hat{f}(x) > 0$, hence the predicted risk premium is never negative. Also, the different risk factors are combined multiplicatively which has shown to be a good model in insurance applications.

Besides for the advantages of GLM with an exponential link, fitting a GLM is relatively fast, the parameters are easily tested for statistical significance and the importance of different explanatory variables in a model is easily analyzed. Furthermore, adding more explanatory variables to the GLM does not significantly change the time to convergence for the parameter estimation algorithm (Dugas et al. 2003). A disadvantage of GLM is that interactions between explanatory variables are not captured in the model, unless these are explicitly defined by means of interaction terms.

2.2.1 Variance function

The variance of Y can be written as

$$\text{var}(Y) = b''(\theta)\ a(\phi) \ , \tag{6}$$

where a and b are as in (5) and $b''(\theta)$ is known as the variance function. Recall that $\theta = \theta(\mu)$. Since $b''(\theta)$ depends on the mean of the distribution of Y, the variance function is often denoted $V(\mu)$.

As an example, the normal distribution has a constant variance function $V(\mu) = 1$, the Poisson distribution has variance function $V(\mu) = \mu$ and the gamma distribution has variance function $V(\mu) = \mu^2$ (McCullagh and Nelder 1983).

2.2.2 Coefficient estimation with Maximum likelihood

The parameters in a GLM are usually estimated with maximum likelihood (Olsson 2002). The log likelihood estimator is on the form

$$l = \sum_i \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \ . \tag{7}$$

In order to maximize (7), the derivative w.r.t. $\beta_j,\ j = 1, \ldots, p$ is taken

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{\mathrm{d}\theta}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\frac{\partial \eta}{\partial \beta_j} \ . \tag{8}$$

We have that $\partial\eta/\partial\beta_j = x_j$, $b'(\theta) = \mu$ and $b''(\theta) = V$ which gives

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{y_i - \mu}{a(\phi)}\frac{1}{V}\frac{\mathrm{d}\mu}{\mathrm{d}\eta}x_j = \left\{W^{-1} = \left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)^2 V\right\} \tag{9}$$

$$= \sum_i \frac{W}{a(\theta)}(y_i - \mu)\frac{\mathrm{d}\eta}{\mathrm{d}\mu}x_j \tag{10}$$

and hence the optimal $\beta_j,\ j = 1, \ldots, p$ are found by solving

$$\sum_i \frac{W_i(y_i - \mu_i)}{a(\phi)}\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}x_{ij} = 0 \ , \tag{11}$$

where $\mu_i = \mu_i(\beta_j)$.

2.2.3 Poisson-Gamma model with frequency and severity

A common procedure for predicting the risk premium is to fit a GLM with claim frequency as dependent variable and another GLM with claim severity as dependent variable. The predictions from each model are then multiplied.

The two main arguments for modeling the risk premium in this way are that (1) claim frequencies are often estimated more accurately and have a larger impact on the resulting risk premium and (2) modeling claim frequencies and severities separately provides more understanding of the resulting risk premium model (Ohlsson and Johansson 2010).

A Poisson distributed GLM with logarithmic link function is often chosen for the claim frequency [claims/year] (Anderson et al. 2004). The Poisson distribution has probability mass function

$$f(y, \mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{12}$$

and variance function $V(\mu) = \mu$.

As for modeling claim severity, a gamma distributed GLM with logarithmic link function is often chosen (Anderson et al. 2004). The gamma distribution has probability density function

$$f(y, \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \ , \tag{13}$$

where the gamma function $\Gamma(\alpha)$ is defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \ . \tag{14}$$

The variance function is $V(\mu) = \mu^2$. Since the range of the gamma distribution is $(0, +\infty)$, a GLM must be fitted for claim sizes > 0.

2.2.4 Tweedie's compound Poisson model

Another common method for modeling the risk premium is to fit a Tweedie distributed log-link GLM with risk premium as dependent variable.

Tweedie distributions are distributions which have a variance function on the form

$$V(\mu) = \mu^d \ . \tag{15}$$

Hence, Tweedie distributions include e.g. the normal distribution (for $d = 0$), the Poisson distribution (for $d = 1$), the gamma distribution (for $d = 2$) and the inverse normal distribution (for $d = 3$) (Ohlsson and Johansson 2010).

Tweedie distributions for which $1 < d < 2$ are compound Poisson distributions which follow the distribution of a Poisson sum of gamma distributed random variables. This distribution has a point mass at zero. For values of d near 1 the distribution resembles a Poisson distribution, and for d near 2 the distribution resembles a gamma distribution (Anderson et al. 2004). Since experience has proven that the Poisson and gamma distributions are suitable for modeling claim frequencies and claim severities respectively, the Tweedie distribution is a suitable choice for modeling the risk premium directly instead of modeling frequencies and severities apart (Ohlsson and Johansson 2010).

A compound Poisson distribution is defined as follows. Assume that N is a Poisson distributed random variable with mean μ, i.e. $N \in \text{Po}(\mu)$. Also assume that $X_1, X_2, \ldots$ are independent identically distributed random variables. Define S_N as

$$S_N = \begin{cases} 0 & \text{if } N = 0 \\ X_1 + \ldots + X_m & \text{if } N = m \geq 1 \end{cases}. \tag{16}$$

Then S_N has a compound Poisson distribution (Haigh 2013). For Tweedie's compound Poisson distribution, the random variables $X_1, X_2, \ldots$ are i.i.d. gamma distributed and the probability function is

$$\begin{cases} f_Y(y;\theta,\lambda,\alpha) = \sum_{n=1}^{\infty} \frac{\{(\lambda\omega)^{1-\alpha}\kappa_\alpha(-1/y)\}^n}{\Gamma(-n\alpha)n!y} \exp\{\lambda\omega(\theta_0 y - \kappa_\alpha(\theta_0))\}, \quad y > 0 \\ p(Y=0) = \exp\{-\lambda\omega\kappa_\alpha(\theta_0)\} \end{cases}, \tag{17}$$

where $\kappa_\alpha(\theta) = (\theta/(\alpha-1)^\alpha((\alpha-1)/\alpha)$, $\theta_0 = \theta\lambda^{1/(1-\alpha)}$ and ω is the exposure (Anderson et al. 2004).

2.3 Artificial neural networks

Artificial neural networks, ANN, form a family of models said to be inspired by the construction of the human brain. An ANN model is formed by a set of computing units, or neurons, connected in various ways and thus forming a network. The network is often structured in layers. Typically, a neural network has an input layer, one or several hidden layers and an output layer. Each neuron receives input data in form of weighted sums of outputs from other neurons, onto which a transform, or activation function, is applied. The result is outputted from the neuron. A network where the signals are only allowed to be transferred from one layer to the next, i.e. forward,

are called feed-forward networks. By varying the network architecture and using different nonlinear activation functions, an ANN can model different types of nonlinear dependencies. The usage of artificial neural networks is among others nonlinear regression, pattern recognition/classification and data clustering (Silva et al. 2017).

2.3.1 Universal approximation theorem

In 1989, Cybenko (Cybenko 1989) showed that any real continuous bounded multivariate function can be approximated by a single layer feed-forward ANN with a sigmoidal activation function. The approximation is on the form

$$G(x) = \sum_{j=1}^{U} \alpha_j \sigma(w_j^T x + \theta_j) \ , \tag{18}$$

where $x \in \mathcal{R}^u, w_j \in \mathcal{R}^u$ and $\alpha_j, \theta_j, U \in \mathcal{R}$. A sigmoidal function $\sigma(t)$ is defined as a function for which it holds that

$$\sigma(t) \to \begin{cases} 1, \ t \to +\infty \\ 0, \ t \to -\infty \end{cases} \ . \tag{19}$$

An example of a sigmoidal function is the sigmoid function

$$r(t) = \frac{1}{1 + e^{-t}} \ . \tag{20}$$

Hence, Cybenko showed that finite sums $G(x)$ are dense in $C(I_u)$, where $C(I_u)$ denotes the space of continuous functions on the u-dimensional hypercube $[0, 1]^u$. This means that that for any $\epsilon > 0$ and function $f \in C(I_u)$, there exists a $G(x)$ s.t.

$$|G(x) - f(x)| < \epsilon \qquad \forall x \in I_u \ . \tag{21}$$

2.3.2 Multilayer perceptron

The multilayer perceptron, MLP, is a commonly used type of feed-forward ANN. An MLP has an input layer with as many neurons as the number of explanatory variables. The input layer is followed by one or several hidden layers with an optional number of neurons. The number of neurons in the output layer equals the number of dependent variables, i.e. in the case of multivariate nonlinear regression the number of output neurons is more than one. Each neuron has an assigned transformation function $r(t)$. Common choices of $r(t)$ are seen in Table 1.

Name	$r(t)$
hyperbolic tangent	$\tanh(t) = (e^{2t} - 1)/(e^{2t} + 1)$
sigmoid	$1/(1 + e^{-t}) = (\tanh(t/2) + 1)/2$
Gaussian	$e^{-t^2/2}$
identity	t
threshold	$\begin{cases} 0, \text{ if } t < 0 \\ 1, \text{ otherwise} \end{cases}$

Table 1: Common choices of activation functions for hidden and output layers in an MLP

Hence, the output from hidden or output layer neuron j is on the form

$$o_j = r(\theta_j + \sum_{i=1}^{l} w_{i,j} o_i) \ , \tag{22}$$

where $\theta_j \in \mathcal{R}$ denotes the bias of neuron j, l is the number of neurons in the previous layer, $w_{i,j} \in \mathcal{R}$ is the weight from neuron i in the previous layer to the neuron j, and o_i is the output from neuron i. The general structure of an MLP is seen in Figure 1. The complexity and flexibility of an MLP is changed by varying the number of hidden layers, activation functions and number of neurons (Sarle 1994).

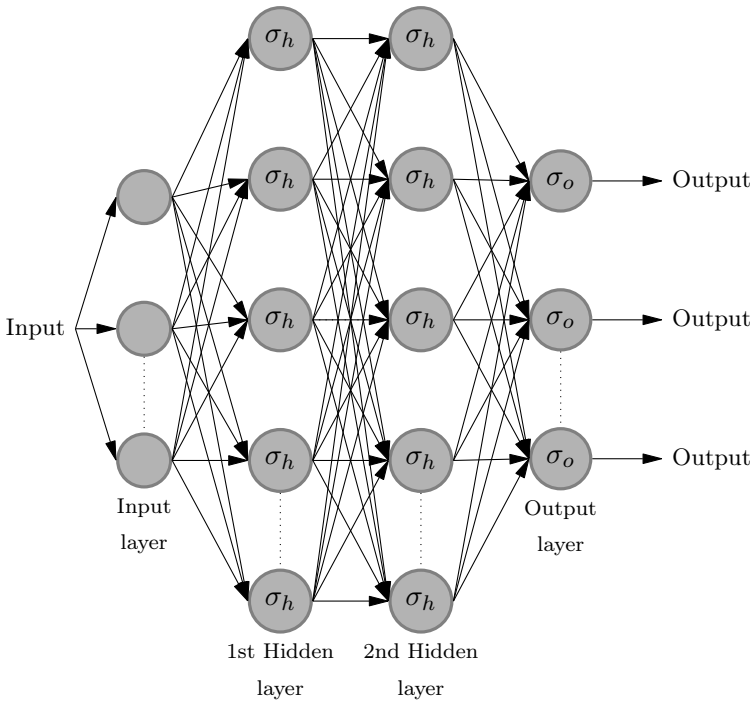

Figure 1: General structure of a multilayer perceptron with 2 hidden layers. In regression, the number of neurons in the input layer equals the number of explanatory variables. Note that different activation functions can be used for the hidden and output layers. The number of neurons in the output layer equals the desired number of outputs. This is a feed-forward ANN where information flows forward in the network. The inputs to the neurons in the hidden and output layers are weighed sums of the outputs from the previous layer. Note that the hidden and output layer neurons usually have a bias term which is added to the weighted sum of inputs.

2.3.3 Back propagation algorithm

The weights in an MLP network are learned using backpropagation, which is an iterative 2-stage supervised learning algorithm. The weights are usually initialized by randomization. In the forward propagation stage, the training data is inputted to the network. The output from the model is compared to the corresponding observed value of the dependent variable. In the back-propagation stage, the weights are adjusted so that the error on the training set is lowered (Silva et al. 2017). This is done by calculating the gradient of

a predefined loss function L,

$$\Delta \mathbf{W} = \frac{\partial L}{\partial \mathbf{W}} \ , \tag{23}$$

where $\mathbf{W}$ is a matrix with the weight and bias parameters of the network (Du and Swamy 2014). The loss function L is typically chosen as the mean squared error. The length of the step taken in the backpropagation stage is controlled by the step size η, which is also known as the learning rate.

2.3.4 Stochastic gradient descent

Stochastic gradient descent is an optimization procedure where the gradient of the loss function L in (23) is not calculated for the whole training set in each iteration. Instead, (23) is calculated for only one observation and then a step is taken in the negative direction of the gradient, i.e. in the direction where the loss function decreases the most (Hastie et al. 2009). This procedure is not as computationally demanding as calculating the gradient for all training observations before taking a step. Hence, using stochastic gradient descent makes the learning process of an MLP faster. Often stochastic gradient descent is implemented with more than one training observation per step. The gradient is calculated for a batch of the larger training set before a step is taken in the negative direction of the gradient. This method is less sensitive to noise in a single observation. The algorithm converges when a predefined tolerance level is reached.

2.4 Model assessment

In this section, the methods chosen for the model comparison are presented.

2.4.1 Mean squared error

The mean squared error, MSE, is defined as

$$\frac{1}{n} \sum_{\{x_i, y_i\} \in \mathcal{S}} (\hat{f}(x_i) - y_i)^2 \ , \tag{24}$$

where $\mathcal{S}$ denotes a data set comprising n observations and $\hat{f}(x_i)$ is the predicted value of y_i given a predictor $\hat{f}(\cdot)$.

In insurance rate making it is necessary that the model is as precise as possible. This is known as the precision criterion (Dugas et al. 2003). When

choosing from a range of candidate models, the model selected should then be the one that minimizes

$$E_{A,X}[(\hat{f}(X) - A)^2] \ . \tag{25}$$

The precision criterion is hence acknowledged when choosing a predictor $\hat{f}(X)$ that minimizes the expected squared error. The true distribution $f(X, A)$ is not known and hence (25) is estimated by the MSE (24). The MSE is an unbiased estimator of the expected squared error on a test set $\mathcal{S}_{test}$, providing that $\mathcal{S}_{test}$ has not been used for fitting the predictor $\hat{f}(X)$ (Dugas et al. 2003).

The expected MSE obtains its minimum for $\hat{f}(X) = E[A|X]$. Indeed, using the tower property, we have that

$$E[\hat{f}(X) - A)^2] = E[(\hat{f}(X) - E[A|X])^2] \ , \tag{26}$$

which is minimized for $\hat{f}(X) = E[A|X]$.

The squared bias of $\hat{f}$ is defined as $(E[A|X] - E[\hat{f}(X)])^2$, where the expectation of $\hat{f}(X)$ refers to the average predictor fitted from the data set at hand. The variance of $\hat{f}(X)$ is defined as $E[(\hat{f}(X) - E[\hat{f}(X)])^2]$. Using these two definitions, we can write the ESE as

$$E[(A - \hat{f}(X))^2] = E[(E[A|X] - E[\hat{f}(X)])^2)] + E[(E[\hat{f}(X)] - \hat{f}(X))^2] + \text{error} \ . \tag{27}$$

This means that the sum of the variance and the squared bias is minimized by choosing a predictor which minimizes the MSE on a test set $\mathcal{S}_{test}$ (Dugas et al. 2003).

2.4.2 Cross-validation

The k-fold cross-validation estimate of the prediction error is defined as (Hastie et al. 2009)

$$\text{CV}(\hat{f}, w) = \frac{1}{k} \sum_{j=1}^{k} L(y_j, \hat{f}^{-j}(x_j, w)) \ , \tag{28}$$

where $\hat{f}^{-j}(x_j, w)$ denotes the prediction from a model with parameters w fitted with the j:th fold, $j = 1, \ldots, k$ removed. Here, x_j denotes the values of the explanatory variables in fold j and y_j are the observed claim amounts in validation set j. The loss function $L(\cdot)$ is often taken as the MSE.

2.4.3 Bootstrap

A method for estimating the variance in prediction error of a predictor is bootstrap. Assume there is a model for which the prediction error is to be tested, a training set comprising n observations and a test set of unseen observations. With bootstrap, the training set is replicated by drawing with replacement n times. The model is then fitted to the replicated training set. The test set is used for prediction and the prediction error, e.g. the mean squared error, is calculated. This procedure is repeated K times, generating a set of prediction errors for different replications of the original training set.

If $\bar{B}$ denotes the bootstrap average of the prediction error and B_l denotes the prediction error for the model fit on the lth, $l = 1, \ldots, K$ bootstrapped training set, then the bootstrap variance is calculated as (Hastie et al. 2009)

$$\widehat{Var}_B = \frac{1}{K-1} \sum_{l=1}^{K} (B_l - \bar{B})^2 \ . \tag{29}$$

2.4.4 Risk ratios

A risk ratio is quotient on the form

$$\frac{\text{claims cost}}{\text{risk premium}} \ , \tag{30}$$

and can be used to evaluate the precision and fairness of a risk premium model. When choosing among a range of candidate models, choosing a *fair* model means favoring models which do not systematically discriminate any subgroup of customers (Dugas et al. 2003).

In a perfectly precise model, (30) equals to 1 for all subsets of policyholders. The fairness criterion of a risk premium model can be addressed by studying the variance of the risk ratios for the variable groupings of the explanatory variables.

2.4.5 Sensitivity

Calculating the sensitivity is a method for analyzing the relative importance of the explanatory variables in a model. The sensitivity of an explanatory variable is the decrease in prediction error of the full model, compared to

the prediction error when that variable is held constant.

First, the explanatory variable for which the sensitivity is to be calculated is set to a constant value. Then, predictions are obtained using the fitted model. Finally, the decrease in prediction error for the full model as a percentage of the prediction error of the model with the specific explanatory variable held constant is calculated. A high sensitivity corresponds to a higher importance of that specific explanatory variable (Francis 2001).

3 Method

This section starts with a presentation of the data preprocessing in terms of choice of explanatory variables and their corresponding grouping, analysis of the claim size distribution and the possible consequences thereof and division of the data into training, validation and test sets. Then follows a presentation of the method for selecting the Tweedie GLM, the Poisson-Gamma GLM and the MLP model.

3.1 Data preprocessing

3.1.1 Explanatory variables

The explanatory variables chosen to be included in the study are *age* [years], *driving distance* per year [km/year], *engine power* [kW], length of *car ownership* [years], *car age* [years], time since receiving *driving license* [years], *population density* at place of residence [people/km^2], whether or not the car is *imported* and *car brand*. These variables are often included in models for the risk premium since they often show good correlation with either the risk premium, the claim frequency or the claim severity. According to actuarial praxis, the continuous variables were grouped before analysis. The grouping was chosen as seen in Table 2. In the case of missing data, the observation was placed in a separate group for the corresponding explanatory variable.

The same explanatory variables with the same grouping were used for prediction on the test set with the final models. Hence, no variable selection methods were used. This is in line with the objectives of this thesis, by which it is necessary to make the comparison between the models' ability of finding patterns in the data as fair as possible. Naturally, in rate making for practical use, variable selection methods ought to be used and different

groupings of the explanatory variables should be tested in order to produce a model with as good predictive properties as possible. This includes deeper analysis of significance of the explanatory variables.

Variable	Unit	Grouping
age	years	< 30, $30 - 44$, $45 - 59$, $60 - 74$, ≥ 75
driving distance	10 km	$0 - 999$, $1000 - 1999$, $2000 - 2999$ $3000 - 3999$, $4000 - 4999$, ≥ 5000 missing data
engine power	kW	$0 - 99$, $100 - 199$, $200 - 299$, $300 - 399$ $400 - 499$, ≥ 500 missing data
car ownership	years	$0 - 4$, $5 - 9$, $10 - 19$, ≥ 20 missing data
car age	years	$0 - 4$, $5 - 9$, $10 - 19$, ≥ 20 missing data
driving license	years	$0 - 9$, $10 - 19$, ≥ 20 missing data
population density	people/km^2	$0 - 999$, $1000 - 1999$, $2000 - 2999$ $3000 - 3999$, ≥ 4000 missing data
imported	-	yes, no missing data
car brand	-	each car brand marks its own group, except for brands with less than 1000 observations which are placed in a separate group missing data

Table 2: Grouping of explanatory variables

3.1.2 Claim size distribution

The raw data set contains roughly $7 \cdot 10^6$ rows, where each row represents a policyholder with a corresponding set of explanatory variables. The data was aggregated on these variables with grouping as in Section 3.1.1. Aggregated rows with a negative sum of claim amounts were removed since these are due to faulty data. The size of the aggregated data set was approximately 10^5 rows.

Notable about the distribution of claim sizes is that it is asymmetric with nearly all of the mass concentrated at zero. The distribution also has a heavy right tail. This means that most of the policyholders do not report any claims while a few policyholders report very large claims. The random occurrence of large claims in training, validation and test sets will largely affect the prediction error for the fitted models. A large claim representing a significant portion of the total claim amount will thus affect the prediction error on the fitted models to a large extent. Such effects will tend to override patterns in the data successfully modeled by one or several of the models. Therefore, these effects need to be limited since they make comparison between the different models more difficult. Also, large claims are probably due to a high degree of randomness for which the GLM and MLP models proposed in this thesis are not suited.

Hence, in order to limit the effects of large claims, the claim sizes were capped at the 99% quantile of the claim amounts > 0. The 99% quantile of the claim amounts was found to be about $1.8 \cdot 10^5$ SEK. Note that the 1% largest claims correspond to 43% of the total claim amount. In Figure 2, a histogram for the capped claim size per policyholder and year is plotted. Figure 3 shows a histogram for the capped log of claims > 0 per policyholder and year. Note the peak to the right in both figures corresponding to the capped claims.

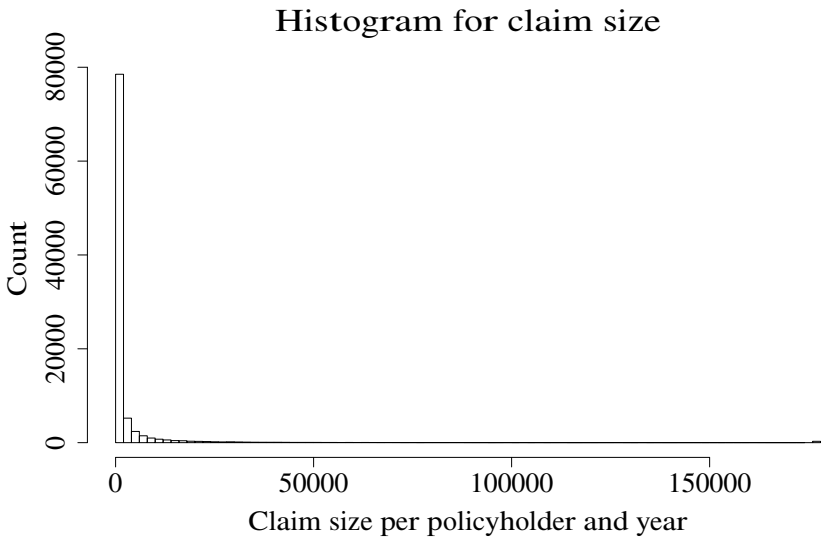

Figure 2: Number of observations per incurred claim size

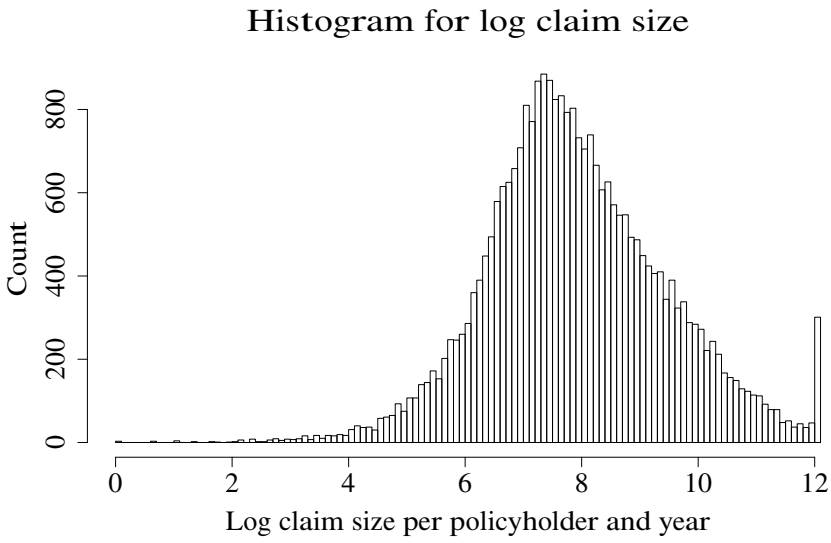

Figure 3: Number of observations per incurred log claim size

3.1.3 Training, validation and test sets

A test set comprising 15% of the data was set aside to be used for final model evaluation. The remaining 85% was used for training and validation.

3.2 Generalized linear models

In this thesis, the GLM modeling is done with two different approaches. The first approach is to model the risk premium as dependent variable. The second approach is to model claim frequency and claim severity separately and then multiply predictions from the claim frequency and claim severity models to obtain predictions of the risk premium. These two GLM modeling approaches are presented in this section.

3.2.1 Tweedie GLM

A logarithmic link GLM with Tweedie's compound Poisson distribution is chosen for modeling the risk premium as dependent variable. The explanatory variables are as described in Section 3.1.1. The specific Tweedie distribution is chosen by 10-fold cross-validation of the variance function parameter d with the remaining coefficients estimated by maximum likelihood. Recall that for Tweedie's compound Poisson distribution, d needs to be chosen in the interval $(1, 2)$. The values of d selected for cross-validation are 9 equidistant points in this interval,

$$d = \{1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9\} \ .$$

The fitted model which yields the lowest cross-validation estimate of the prediction error is chosen as the final model.

3.2.2 Poisson-Gamma GLM

The other approach within the GLM framework is modeling claim frequency and claim severity separately. The predictions from each model are then multiplied to form predictions of the risk premium, since it holds that

$$\text{risk premium} = \text{claim frequency} \cdot \text{claim severity} \ , \tag{31}$$

where claim frequency is the number of claims per year and claim severity is the total claim amount divided by the number of claims.

Claim frequency is assumed to follow a Poisson distribution and claim severity is assumed to be gamma distributed, according to actuarial praxis. Since

the domain of definition for the gamma distribution is $x > 0$, a logarithmic link gamma GLM with claim severity as dependent variable is fitted to the subset of the training data with claim severity > 0. For the frequency model, a logarithmic link Poisson GLM is fitted to the training data with frequency as dependent variable. The explanatory variables and corresponding grouping are as in Section 3.1.1.

3.3 Multilayer perceptron models

In this section, the process of choosing an MLP model is described.

3.3.1 Choice of activation functions

Choosing a sigmoidal activation function for the MLP is motivated based on the universal approximation theorem, presented in Section 2.3.1. Furthermore, since the risk premium is non-negative, it is desired that the range of the activation function for the output layer does not cover any part of $(-\infty, 0)$. The sigmoid function,

$$r(t) = \frac{1}{1 + e^{-t}} \ , \tag{32}$$

is a common choice of activation function for an MLP. It has range $(0, 1)$, which suits the positivity requirement. It is also known for giving a generally good performance in MLP regression models. Hence, the sigmoid function is chosen as activation function for the hidden and output layers.

3.3.2 Network architecture

The space of possible network architectures is infinite. It is therefore necessary to choose a subset of network configurations and test their performance on the data set at hand. Choosing an MLP with one hidden layer is supported by the universal approximation theorem, see Section 2.3.1. Some initial tests indicate that a shallow MLP produces the best results in terms of training MSE and convergence time, while networks with a larger number of hidden layers have shown poor convergence in initial tests.

The subset $\mathcal{M}$ of MLP models selected for further investigation are

$$\begin{aligned}\mathcal{M} = \{&[9, 5, 1], [9, 10, 1], [9, 20, 1], [9, 30, 1], [9, 40, 1], \\ &[9, 50, 1], [9, 10, 10, 1], [9, 30, 10, 1], [9, 50, 10, 1]\} \ ,\end{aligned}$$

where $[x, y, z]$ denotes a network architecture with x, y and z neurons in the input, hidden and output layer(s). As seen, there are both 1- and 2-hidden layer networks among the network architectures to be tested further. The number of input neurons is fixed at the number of explanatory variables. Similarly, there is one output neuron since the desired number of output values is one: the predicted risk premium. Hence, the variable parameters in the network architecture are the number of hidden layers and the number of neurons in each hidden layer. It is necessary to adjust the number of parameters in the network to the data set in order to avoid overfitting. A network with too many degrees of freedom easily overfits the training data by modeling noise. Such a model will thus not perform well on tests with previously unseen data.

The number of neurons in the first hidden layer among the selected architectures are chosen to span a large range. Similar values have also shown promising results in initial tests. For the models with a second hidden layer, the number of second hidden layer neurons is chosen to 10. This is for comparability as well as limiting the number of degrees of freedom to avoid overfitting.

For each model in $\mathcal{M}$, a 5-fold cross-validation is performed. The models are trained using backpropagation with stochastic gradient descent. Also, the claim sizes in the training data are normalized to be in the range of the sigmoid function. Hence, for an observation x_i, the maximum value $x^{\max}$ and the minimum value $x^{\min}$ in the training set, the normalized x_i^{norm} is calculated as

$$x_i^{\text{norm}} = \frac{x_i - x^{\min}}{x^{\max} - x^{\min}} \ . \tag{33}$$

As seen, x_i^{norm} has range $[0, 1]$. The normalized predictions are transformed back to the original range of the claim sizes with the inverse of (33).

The cross-validation error is calculated for each model, as well as the average time required until convergence and the corresponding average number of epochs. A well-performing model should both have a low cross-validated estimate of the prediction error as well as converge in a reasonable amount of time and number of iterations.

For this first stage in the model selection process, a tolerance level of `tol_level` $= 0.003$ is chosen. This proved to be a reasonably good value during initial testing. The batch size is chosen to `batch_size` $= 12000$, which

corresponds to about 20% of the number of observations in the training set. The learning rate is set to `learn_rate` $= 0.1$. These initial values for the batch size and learning rate also proved to be reasonable values during the initial testing. The rationale for why the model architecture and model parameters are cross-validated separately is shortage of time given the available computing power. Cross-validating the model architecture in combination with the model parameters would have been more optimal.

After completing the 5-fold cross-validation, the model performing the best based on the measures described above is selected for further parameter calibration. For the selected model, the tolerance level, batch size and learning rate are cross-validated according to the following scheme:

- `tol_level` $= [0.0029,\ 0.0030,\ 0.0031]$ with `batch_size` $= 12000$ and `learn_rate` $= 0.1$
- `batch_size`$\times$`learn_rate` $= [0.01,\ 0.05,\ 0.1]\times[3000,\ 6000,\ 12000,\ 18000]$ with tolerance level set to the best-performing tolerance level from the previous cross-validation.

In total, 5-fold cross-validation is performed for $(3 + (3 \times 4)) = 15$ models. Thus, in total 75 models are fitted for the parameter selection.

Cross-validating the tolerance level, batch size and learning rate together was considered too time-consuming, providing that a reasonable number of values for each parameter should be cross-validated. If e.g. three values for each parameter is to be cross-validated, the number of models to fit is $5 \cdot 3^3 = 135$. Given an average time to convergence for each fit of 5 hours, the cross-validation of parameters would take a month in total.

Therefore, the tolerance level was chosen to be cross-validated separately from the batch size and learning rate. The reason for this particular setting is that the choice of batch size and learning rate are parameters affecting the learning process of the network, while the choice of tolerance level has a more statistical meaning since it affects over- and underfitting the training data and is thus related to the bias-variance trade off. If the tolerance level is set too low, the network will most likely overfit the training data, yielding high prediction errors on unseen data. If the tolerance level is set too generously, the network will underfit the training data and thus not learn the general structure of the data. Hence, the prediction error on unseen test data will be poor.

The tolerance level and combination of batch size and learning rate yielding the lowest cross-validated estimate of the prediction error as well as a reasonable average time to convergence and corresponding number of iterations are the parameters selected for the final MLP model. This concludes the selection of MLP model.

4 Results

This section comprises the results of the study and starts with the results from the model selection process. Then follows the results from the model comparison with the different measures presented in Section 2.4.

4.1 Cross-validation of MLP models

The results from the 5-fold cross-validation of the candidate MLP models in $\mathcal{M}$ are shown in Table 3. The table shows the cross-validated estimate of the prediction MSE denoted CV error, the average time until convergence and the average number of epochs required for convergence.

Model	CV error	Time [h]	Epochs
$[9, 5, 1]$	$1.8899 \cdot 10^8$	1.7	567 000
$[9, 10, 1]$	$1.8904 \cdot 10^8$	2.1	478 000
$[9, 20, 1]$	$1.8899 \cdot 10^8$	3.0	443 000
$[9, 30, 1]$	$1.8897 \cdot 10^8$	3.3	344 000
$[9, 40, 1]$	$1.8899 \cdot 10^8$	4.2	353 000
$[9, 50, 1]$	$1.8903 \cdot 10^8$	4.7	317 000
$[9, 10, 10, 1]$	$1.8895 \cdot 10^8$	10.9	1 582 000
$[9, 30, 10, 1]$	Not attempted	-	-
$[9, 50, 10, 1]$	Not attempted	-	-

Table 3: Cross-validation results for MLP models

Note that the models $[9, 30, 10, 1]$ and $[9, 50, 10, 1]$ were not attempted due to the poor convergence of $[9, 10, 10, 1]$. Models $[9, 30, 10, 1]$ and $[9, 50, 10, 1]$ are extended versions of $[9, 10, 10, 1]$ and the time it would take for these models to converge is too large given that there needs to be enough time to cross-validate the model parameters within the time scope of this thesis.

Model $[9, 10, 10, 1]$ has the lowest CV error of all the models in $\mathcal{M}$, although

the difference is rather small. However, since $[9, 10, 10, 1]$ took significantly more time to fit than the other models, leaving little time for further parameter cross-validation, and since the improvement in prediction error was not very large, this model was not considered to be a candidate for further calibration.

The model with the second lowest CV error is $[9, 30, 1]$. Compared to the other models it also has a rather low average number of epochs until convergence and an average time until convergence that allows for further parameter adjustments within the time scope of this thesis. Hence, $[9, 30, 1]$ was chosen for further cross-validation of the parameters `tol_level`, `batch_size` and `learn_rate`.

In Table 4 the results from the 5-fold cross-validation of `tol_level` for model $[9, 30, 1]$ are shown. As before, CV error denotes the cross-validated estimate of the prediction MSE, time is the average time for the learning algorithm to converge and epochs is the corresponding number of epochs required. The lowest CV error was obtained for `tol_level` $= 0.0030$, why this is the tolerance level used for cross-validation of `batch_size` and `learn_rate`. The cross-validation results for the different values of `batch_size` and `learn_rate` are found in Table 5. As seen, the lowest cross-validated estimate of the prediction error is obtained for `batch_size` $= 6000$ and `learn_rate` $= 0.05$.

Thus the final MLP model was chosen to be $[9, 30, 1]$ *with* `tol_level` $=$ *0.0030,* `batch_size` $=$ *6000 and* `learn_rate` $=$ *0.05.*

`tol_level`	CV error	Time [h]	Epochs
0.0029	No convergence	-	-
0.0030	$1.8897 \cdot 10^8$	3.3	344 000
0.0031	$1.9504 \cdot 10^8$	0.7	66 000

Table 4: Cross-validation results for different values of `tol_level` with `learn_rate` $= 0.1$ and `batch_size` $= 12000$

batch_size	learn_rate	CV error	Time [h]	Epochs
3000	0.01	No convergence	-	-
3000	0.05	$1.8904 \cdot 10^8$	5.1	754 000
3000	0.1	$1.8903 \cdot 10^8$	3.0	415 000
6000	0.01	No convergence	-	-
6000	0.05	$1.8892 \cdot 10^8$	5.9	760 000
6000	0.1	$1.8902 \cdot 10^8$	3.2	399 000
12000	0.01	No convergence	-	-
12000	0.05	$1.8900 \cdot 10^8$	6.8	747 000
12000	0.1	$1.8897 \cdot 10^8$	3.3	344 000
18000	0.01	No convergence	-	-
18000	0.05	$1.8901 \cdot 10^8$	8.2	775 000
18000	0.1	$1.9071 \cdot 10^8$	3.0	275 000

Table 5: Cross-validation results for different values of `batch_size` and `learn_rate`, with `tol_level` = 0.003

4.2 Cross-validation of Tweedie GLM

The 10-fold cross-validation estimate of the prediction MSE for different choices of distributions among Tweedie's compound Poisson distribution, defined by the choice of d, are seen in Table 6.

d	CV error
1.1	$1.8264 \cdot 10^8$
1.2	$1.8281 \cdot 10^8$
1.3	$1.8301 \cdot 10^8$
1.4	$1.8309 \cdot 10^8$
1.5	$1.8320 \cdot 10^8$
1.6	$1.8336 \cdot 10^8$
1.7	$1.8360 \cdot 10^8$
1.8	No convergence
1.9	No convergence

Table 6: 10-fold cross-validation estimate of the prediction MSE for GLM models with different Tweedie's compound Poisson distributions, defined by the choice of d

From Table 6 it is seen that the GLM with a Tweedie distribution defined by $d = 1.1$ gives the lowest estimate of the prediction error. Note that Tweedie's

compound Poisson distribution with variance function $V(\mu) = \mu^{1.1}$ is very similar to a Poisson distribution, for which $d = 1$. Indeed, when performing a 10-fold cross-validation of a Poisson GLM, the cross-validated estimate of the prediction error is lowered even further. Hence, a pure Poisson GLM seems to be a more appropriate model than a Tweedie GLM given the CV error.

Nevertheless, since it on beforehand has been decided to make the model comparison between an MLP, a Poisson-Gamma GLM and a Tweedie GLM, the log link Tweedie GLM with variance function $V(\mu) = \mu^{1.1}$ is chosen for the final Tweedie model.

4.3 Model comparison

4.3.1 Mean squared error on test set

The three final models were fitted on the full training set and predictions of the risk premium were then obtained using the until now unseen test set. The corresponding prediction errors are seen in Table 7. For comparison, a model assigning the average of the total claim size on the training set, denoted Intercept, is included in Table 7 with the corresponding prediction MSE on the test set. The average claim size on the training set is 2920 SEK, where the average is taken over all policyholders. Similarly, this figure for the test set is 2730 SEK.

As expected, the undifferentiated intercept model performs the worst on the test set. The Tweedie GLM has the lowest test MSE, followed by the MLP model and the Poisson-Gamma GLM. Given the average claim size on the test set, the prediction MSE in Table 7, and even more the prediction RMSE which is rather similar for all models, it is obvious that the prediction error is highly affected by the heavy right tail of the claim distribution, i.e. the existence of a few very large claims which none of the models have been able to predict.

Model	Prediction MSE	Prediction RMSE
MLP	$1.63 \cdot 10^8$	12 770
Tweedie	$1.57 \cdot 10^8$	12 530
Poisson-Gamma	$1.65 \cdot 10^8$	12 850
Intercept	$1.68 \cdot 10^8$	12 960

Table 7: Prediction MSE and RMSE on the test set for the three different models and the reference Intercept model

4.3.2 Aggregated risk ratio

In Table 8, the aggregated risk ratio on the test set for each model is presented. The aggregated risk ratio is calculated as the quotient of the total sum of claims cost and the total sum of predicted risk premiums. Also, a variance measure is presented, calculated as the variance of the risk ratios on all subgroups of the explanatory variables. Given the fairness criterion, the value of this variance measure should preferably be as low as possible.

From Table 8 it is seen that the Tweedie GLM has the aggregated risk ratio closest to one, followed by the MLP model and then the Poisson-Gamma GLM. The Poisson-Gamma GLM has an aggregated risk ratio of 0.76, meaning that on total this model predicts 32 % higher premiums than motivated by the observations in the test set. The Tweedie GLM has the lowest risk ratio variance, closely followed by the Poisson-Gamma GLM. The MLP has a significantly higher risk ratio variance, corresponding to a standard deviation of 0.4. This is compared to a standard deviation of 0.17 for the Tweedie GLM and 0.2 for the Poisson-Gamma GLM.

From a profitability perspective it is satisfactory to see that the aggregated risk ratio for all models is not higher than one. This means that the total sum of claims cost is lower than the total sum of predicted risk premiums. The risk ratio for the Intercept model is simply the quotient of the average claim sizes for the training and the test set.

Measure	MLP	Tweedie	Poisson-Gamma	Intercept
Risk ratio	0.87	0.95	0.76	0.94
Variance	0.16	0.03	0.04	0.10

Table 8: Aggregated risk ratio on test set and variance of risk ratios

4.3.3 Risk ratios on subsets of policyholders

In this section, risk ratios for the explanatory variables with grouping as before are presented. With respect to the precision criterion, a good model has risk ratios close to one for all explanatory variables and corresponding variable groups. The fairness criterion gives that the risk ratios should vary as little as possible. This means that the model does not systematically discriminate a certain group of policyholders. For each explanatory variable, the weighted mean of the risk ratios for each variable group is also presented, as well as the variance of the risk ratios. The reason why the mean for each model usually differs from the aggregated risk ratio in Section 4.3.2 is that the risk ratios for missing values are not presented.

Note that it is seen from the Intercept risk ratios how the risk varies between the variable groups, since the Intercept is simply the quotient of the observed claim cost on the test set and the average claim cost on the training set. Hence, if the Intercept risk ratio is > 1, the corresponding group should have a higher risk premium than average and the opposite for Intercept risk ratios < 0.

In Table 9, risk ratios for the variable *age* are presented. From the Intercept risk ratios, it is seen that the risk decreases with increasing age, except for the oldest policyholders. This result is in accordance with actuarial knowledge. From the risk ratios for the MLP model, it is seen that the model captures the risk well for the three youngest groups, and less well for the groups $60 - 74$ and ≥ 75, which on average pay 43% and 28% more than motivated by the observations from the test set. The Poisson-Gamma GLM performs the worst compared to the MLP and Tweedie GLM on all groups except for ≥ 75. As seen, the variance is low for all models. Based on the low variance and mean close to one, the Tweedie model performs the best given the fairness and precision criteria for this explanatory variable.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
< 30	0.90	0.83	0.65	1.20	1378
$30 - 44$	1.00	0.91	0.71	0.98	3879
$45 - 59$	0.94	1.03	0.82	0.95	3871
$60 - 74$	0.70	0.94	0.76	0.76	3375
≥ 75	0.78	1.05	0.96	0.93	1599
Mean	0.87	0.95	0.76	0.94	
Variance	0.01	0.01	0.01	0.02	

Table 9: Risk ratios for the explanatory variable *age*

In Table 10, risk ratios for the explanatory variable *driving distance* are presented. From the Intercept risk ratios it is seen that the risk increases with a longer driving distance. This is expected, since a longer yearly driving distance increases the exposure to risk. The model with the mean closest to one and lowest variance is the Tweedie GLM, followed by the Poisson-Gamma GLM. Note that the MLP model has a significantly higher variance than the other two models, corresponding to a standard deviation of 0.67. This is largely due to difficulties predicting the risk premium for the shortest and the two longest distance groups.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0 - 999$	0.48	1.02	0.86	0.54	3627
$1000 - 1999$	0.62	0.90	0.82	0.55	4193
$2000 - 2999$	1.10	0.96	0.85	0.78	2147
$3000 - 3999$	1.32	0.80	0.59	0.76	867
$4000 - 4999$	1.94	0.96	0.75	0.97	332
≥ 5000	2.12	0.79	0.62	0.92	254
Mean	0.71	0.93	0.80	0.63	
Variance	0.45	0.01	0.01	0.03	

Table 10: Risk ratios for the explanatory variable *driving distance*

In Table 11, risk ratios for the explanatory variable *driving license* are presented. As seen from the Intercept risk ratios, the risk decreases the longer a policyholder has had his or her driving license. This is also expected, since a longer driving experience should decrease the risk of e.g. accidents. Again, the Tweedie GLM has the lowest variance and an average risk ratio closest to one. The MLP performs better than the Poisson-Gamma GLM on groups $0 - 9$ and $10 - 19$, but worse for the ≥ 20 group, which comprises the

largest number of policyholders. The average risk ratio is 0.94 for the MLP, which is significantly better than the Poisson-Gamma GLM mean of 0.79. The variance is twice as high for the MLP compared to the Poisson-Gamma GLM, meaning that the MLP is a less fair model than the Poisson-Gamma GLM in the *driving license* dimension. However, the risk ratio variance for all models is relatively small.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0-9$	1.21	0.92	0.69	1.33	2313
$10-19$	0.94	0.91	0.73	1.00	3149
≥ 20	0.79	1.09	0.96	0.84	5083
Mean	0.94	0.98	0.79	1.01	
Variance	0.04	0.01	0.02	0.06	

Table 11: Risk ratios for the explanatory variable *driving license*

In Table 12, risk ratios for the explanatory variable *direct import* are presented. The Intercept risk ratios on the test set indicate that policyholders with imported cars should pay a lower risk premium, which was not expected. Note however that there are less observations in the 'yes' group. Again, the Tweedie GLM has the average risk ratio closest to one and a low risk ratio variance, indicating that the Tweedie GLM is the most precise and fair model in this dimension. The variance of the MLP model is marginally lower than for the Tweedie model. The average risk ratio for the MLP is significantly worse compared to the Tweedie GLM, and marginally better than the average risk ratio of the Poisson-Gamma GLM. The Poisson-Gamma GLM has the highest risk ratio variance of all models for this explanatory variable, although it is at a relatively low level.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
no	0.81	0.98	0.84	0.85	11161
yes	0.85	0.85	0.59	0.78	2663
Mean	0.81	0.95	0.78	0.84	
Variance	0.00	0.01	0.03	0.00	

Table 12: Risk ratios for the explanatory variable *direct import*

In Table 13, risk ratios for the explanatory variable *car age* are presented. The risk decreases with increasing car age, as seen from the Intercept risk ratios. This is expected since older cars are often less worth and less tech-

nological, which makes them cheaper to replace or repair. Also, choosing to drive an old car compared to a newer might indicate that the policyholder has a less risk-prone personality.

For the model comparison along this dimension, it is seen in Table 13 that the Tweedie GLM has an average risk ratio closest to one, and risk ratios closest to one for all variable groups except the largest, $5-9$, where the MLP model performs slightly better. The risk ratio variance of the Tweedie GLM is again very low, and compares to the variance of the Poisson-Gamma GLM. Note that the variance of the MLP model is significantly higher than the variance of the other two models, much depending on the poor risk ratio for the group ≥ 20. However, the average risk ratio is better for the MLP model compared to the Poisson-Gamma GLM. Hence, the MLP model is more precise along this dimension, but less fair than the Poisson-Gamma GLM.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0-4$	1.11	0.98	0.78	1.25	4362
$5-9$	1.03	0.96	0.79	1.05	4913
$10-19$	0.57	0.88	0.69	0.62	3989
≥ 20	0.15	0.72	0.60	0.16	838
Mean	0.87	0.95	0.76	0.94	
Variance	0.20	0.01	0.01	0.23	

Table 13: Risk ratios for the explanatory variable *car age*

See Table 14 for risk ratios for the explanatory variable *engine power*. As expected, the risk increases with increasing power. The reason for this is that cars with a high engine power are often more expensive to replace or repair. Also, choosing to drive a car with a high engine power might indicate that the policyholder is more prone to risk.

As seen in Table 14, the Tweedie GLM has risk ratios very close to one for groups $0-99$ and $100-199$, but the MLP model has a slightly more accurate prediction for group ≥ 200. Given the average risk ratio, the Tweedie GLM is the most precise model, followed by the MLP model and then the Poisson-Gamma GLM. The Tweedie GLM also has the lowest variance, followed by the Poisson-Gamma GLM and the MLP model.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0-99$	0.63	1.00	0.95	0.68	5444
$100-199$	0.88	0.98	0.83	0.88	6356
≥ 200	1.16	0.83	0.56	1.16	2009
Mean	0.81	0.95	0.78	0.84	
Variance	0.07	0.01	0.04	0.06	

Table 14: Risk ratios for the explanatory variable *engine power*

In Table 15, risk ratios for the explanatory variable *car ownership* are found. As seen from the Intercept risk ratios, the risk decreases the longer a policyholder has owned their car. Note however that the number of observations in groups $15-19$ and ≥ 20 are comparably few. The Tweedie GLM performs the best in terms of average risk ratio, followed by the Poisson-Gamma GLM. The MLP model has with little margin the lowest variance of all models, but is the worst performing model from a precision perspective.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0-4$	0.72	1.01	0.79	0.78	5226
$5-9$	0.98	0.89	0.75	0.59	3428
$10-14$	0.36	0.81	0.70	0.31	1254
$15-19$	0.28	1.17	1.50	0.21	367
≥ 20	0.03	0.17	0.17	0.02	137
Mean	0.72	0.95	0.78	0.63	
Variance	0.14	0.15	0.22	0.09	

Table 15: Risk ratios for the explanatory variable *car ownership*

See Table 16 for risk ratios for the explanatory variable *population density*. Note that there are no observations in the test set for group $2000-2999$. There are relatively few observations in the group $3000-3999$, and disregarding the Intercept risk ratio for this group, there seems to be a trend towards increasing risk with increasing population density. This is expected, since a higher density should decrease availability on roads and thus increase the risk of incidents.

Again, the Tweedie GLM has the lowest average risk ratio and risk ratio variance for this explanatory variable. The Poisson-Gamma GLM also has a low variance, but performs the worst from a precision perspective. The MLP model has the highest variance, although it is still very low at 0.02.

The average risk ratio is in between the risk ratios of the other two models.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
$0 - 999$	0.88	0.91	0.78	0.87	7116
$1000 - 1999$	0.98	1.02	0.79	1.07	3413
$3000 - 3999$	0.68	0.88	0.72	0.79	712
≥ 4000	0.80	0.96	0.69	1.01	2676
Mean	0.88	0.95	0.76	0.94	
Variance	0.02	0.00	0.00	0.02	

Table 16: Risk ratios for the explanatory variable *population density*

In Table 17, risk ratios for the explanatory variable *car brand* are presented, for the car brands with the largest number of observations. Note however that the number of observations is relatively low. The Tweedie GLM performs the best in terms of the precision criterion as well as the fairness criterion. The MLP model has an average risk ratio significantly better than the Poisson-Gamma GLM. However, the risk ratio variance is slightly higher than the variance of the Poisson-Gamma GLM, although it is at a low level.

Group	MLP	Tweedie	Poisson-Gamma	Intercept	Count
Volvo	0.90	1.08	1.05	0.91	1551
BMW	1.27	1.13	0.71	1.39	940
Mercedes	1.14	0.85	0.64	1.13	907
Ford	0.75	0.99	0.75	0.83	758
SAAB	0.75	1.02	0.82	0.80	704
Audi	1.07	0.91	0.74	1.20	644
Mean	0.98	1.00	0.78	1.04	
Variance	0.05	0.01	0.02	0.05	

Table 17: Risk ratios for the explanatory variable *car brand*

To conclude the analysis of risk ratios on subsets of policyholders, it is clear that the Tweedie GLM is the most precise and fair model given the subsets analyzed. It has an average risk ratio close to one for all subsets observed which means good precision. The model also has a low variance, which indicates low discrimination. For the comparison between the MLP model and the Poisson-Gamma GLM, the MLP model on average proved to be more precise than the Poisson-Gamma GLM in most cases. However, it generally showed a larger risk ratio variance, which means that it is less fair compared

to the Poisson-Gamma GLM.

4.3.4 Risk ratios on magnitudes of claim size

In Table 18, risk ratios on the test set for different magnitudes of claim size are presented. The first line in Table 18 shows risk ratios for customers without claims. These represent 70% of the policyholders in the test set. For customers with claim size larger than zero, risk ratios for quantile intervals of the claim size distribution are presented. Note that the maximum claim size in the test set, 177340 SEK, is a capped value at the 99th quantile of the original claim size distribution, for claim sizes larger than zero.

The risk ratio measure for different claim sizes should be used as a tool for understanding how the distribution of the predictions from the models relate to actual observations, and not how well the models predict the actual risk. The difference in risk between different groups of policyholders must be analyzed by means of explanatory variables, and not by grouping the policyholders by claim sizes. The high degree of randomness in claim occurrences results in that on average high risk policyholders occur in the zero claim group, and on average low risk policyholders occur in the large claim size group. This is why none of the models were successful in predicting either small or large claims.

Quantile	Claim size	MLP	Tweedie	Poisson-Gamma	Intercept
–	0	0.00	0.00	0.00	0.00
$0-10$	$1-456$	0.09	0.17	0.18	0.09
$10-20$	$457-810$	0.22	0.37	0.38	0.22
$20-30$	$811-1214$	0.35	0.51	0.52	0.35
$30-40$	$1215-1684$	0.50	0.71	0.72	0.49
$40-50$	$1685-2366$	0.68	0.83	0.80	0.69
$50-60$	$2367-3464$	0.92	1.03	0.92	0.98
$60-70$	$3465-5439$	1.31	1.40	1.23	1.48
$70-80$	$5440-9587$	1.99	1.95	1.63	2.47
$80-90$	$9588-21359$	3.60	2.98	2.43	4.87
$90-95$	$21360-41607$	6.88	4.51	3.47	10.05
$95-100$	$41608-177340$	17.55	10.76	8.04	32.03

Table 18: Risk ratios for zero claims and risk ratios for quantile intervals of the claim size distribution, for claim sizes > 0, for policyholders from the test set

4.3.5 Bootstrap estimates of variance in MSE

In Table 19, the results from the bootstrap estimates of variance of the models are presented. Due to the large amount of time required to fit the MLP model on each bootstrap resample of the training data, a 5-resample bootstrap was performed for all models. The results indicate that the MLP model has the highest variance in MSE, followed by the Poisson-Gamma GLM. Note also that not only did the MLP have the highest bootstrap estimate of variance in MSE, the training algorithm of the 5th bootstrap resample did not converge. It must be kept in mind that the MLP model and its parameters have been chosen based on the original training set. The conclusion from the bootstrapping is therefore that the configuration and parameters of an MLP is sensitive to changes in training data.

Model	Bootstrap variance in MSE
MLP*	$5.5 \cdot 10^{13}$
Tweedie	$8.1 \cdot 10^{10}$
Poisson-Gamma	$1.5 \cdot 10^{13}$
Intercept	$6.8 \cdot 10^{8}$

Table 19: Bootstrap estimates of variance in MSE calculated with 5 bootstrapped samples

4.3.6 Sensitivity of explanatory variables

In Table 20, sensitivities for all explanatory variables are shown. The sensitivity has been calculated by first fitting the models on the training data. Then an explanatory variable is set to its most frequent value and predictions are obtained on the training set. The sensitivity is then calculated as the percentage change in prediction MSE for the full model, compared to the model with one explanatory variable set to a constant value.

The sensitivity measure is a way of understanding which explanatory variables are assigned the greatest importance in explaining the risk premium. As seen in Table 20, *driving distance* is the explanatory variable with the highest sensitivity for both the MLP and the Tweedie GLM, and the second highest for the Poisson-Gamma GLM. The MLP and the Tweedie GLM also share the explanatory variable with the second highest sensitivity, which is the length of *car ownership*. The MLP sensitivities for the remaining variables are low. The Tweedie GLM assigns a fairly high significance for variables *car age* and *engine power*. Note that except for *car age* and *driv-*

ing distance, all the sensitivities for the Poisson-Gamma GLM are negative. Hence, in removing a degree of freedom by setting an explanatory variable to a constant value, the prediction MSE on the training set is lowered. This is not entirely surprising, since predictions from the Poisson-Gamma model are products of predictions from two separate models. Indeed, the sensitivity results indicate that there is room for improvement of the Poisson-Gamma model.

Explanatory variable	MLP	Tweedie GLM	P-G GLM	Constant
driving distance	3.50	4.01	0.36	$1000 - 1999$
car ownership	3.00	2.78	-0.86	$0 - 4$
age	0.20	0.15	-0.48	$30 - 44$
direct import	0.16	0.20	-1.95	no
engine power	0.12	0.80	-2.60	$100 - 199$
population density	0.12	0.00	-2.22	$0 - 999$
car brand	0.09	0.22	-2.48	other
car age	0.05	0.95	2.04	$5 - 9$
driving license	0.02	0.06	-2.47	≥ 20

Table 20: Sensitivity [%] on the training set for each explanatory variable and the corresponding constant value of the explanatory variable. The explanatory variables are ordered by decreasing sensitivity in the MLP model, i.e. by decreasing order of relative importance.

5 Discussion

5.1 Return to objectives

The objective of this thesis is to analyze how multilayer perceptron modeling compares to traditional GLM modeling for pricing car damage insurance. MLP models have an advantage to GLM in that they, depending on their construction, have the ability to model high-dimensional nonlinear interaction effects between explanatory variables. Interaction effects are very likely to exist in car damage insurance data. With GLM, such interaction terms need to be manually found and integrated in the model. Also, in contrast to an MLP model a GLM is constrained by the assigned distribution of the outputs. Given these two arguments, it should be possible to construct an MLP which is preferable to GLM, by means of actuarial precision and fairness.

As for model precision, the Tweedie GLM clearly shows highest precision of the three models. First, the Tweedie GLM has the lowest prediction MSE on the test set, followed by the MLP model and last the Poisson-Gamma GLM. Concerning model precision from the perspective of each explanatory variable, the Tweedie GLM proves to be the most precise in each dimension. The MLP model is more precise than the Poisson-Gamma GLM for 7 of 9 explanatory variables.

The precision results from the MLP model are promising for the use of artificial neural networks in car damage insurance rate making, given that the chosen MLP configuration in this thesis is fairly simple. The Tweedie and Poisson-Gamma GLM approaches are results from statistical assumptions of underlying processes behind the data based on actuarial understanding, whereas the MLP model is free from previous assumptions. Given the good precision of the MLP model, it can be concluded that the it succeeds in modeling important structures in the data. It can also be concluded that the distributional assumptions for the GLM models, especially the Tweedie GLM, suited the data well.

The other interesting aspect of the precision results is that the Tweedie model performs significantly better than the Poisson-Gamma GLM. Naturally, the models in this thesis need to be further calibrated before practical use, in terms of variable selection and grouping. This especially concerns the Poisson-Gamma GLM, which had a negative sensitivity for a majority of the explanatory variables. However, the results show that a Tweedie GLM is more suitable than a Poisson-Gamma GLM for this particular data set, in terms of modeling structural behavior.

Regarding the fairness criterion, the results show that not only is the Tweedie GLM the most precise model, but also the most fair. The Tweedie GLM has the lowest total risk ratio variance and a low risk ratio variance for all explanatory variables. The comparison between the MLP model and the Poisson-Gamma GLM shows that the MLP has a four times higher aggregated risk ratio variance than the latter. Given the subgroups analyzed, this indicates that the MLP model is less fair than the Poisson-Gamma GLM.

Indeed, the risk ratio variance both in total and for a majority of the explanatory variables is the highest for the MLP model, but the results are still comparable to the other two models. What is more worrying is the results from the bootstrapping of the training set. The MLP model has the highest

bootstrapped variance in MSE. Taking the fourth root of this figure gives a standard deviation of the risk premium prediction of 2720 SEK, compared to 1970 SEK for the Poisson-Gamma GLM and 530 SEK for the Tweedie GLM. These results, in combination with the fact that the MLP model did not converge for one of the bootstrap resamples, show that the MLP is a less stable model than the other two.

That the Tweedie GLM is more fair than the Poisson-Gamma GLM for this data set is interesting, especially in combination with the fact that the Tweedie GLM proved to be more precise. From a perspective of precision and fairness, the Tweedie GLM has hence better performance and is more suitable for modeling the risk premium than the Poisson-Gamma GLM, for this particular data set.

5.2 Practical implementation

From the perspective of practical implementation in a real-world insurance business, rate making with artificial neural networks has two main advantages compared to rate making with GLM.

The first practical advantage is that with artificial neural networks there is no need to spend time on finding interaction effects between explanatory variables, as these will automatically be modeled in a well-configured network. Finding interaction effects manually, especially high-dimensional, can prove to be difficult and time consuming. Hence, such effects are usually neglected with GLM, which in turn introduces bias to the model.

The other main practical advantage of using artificial neural networks in rate making applications should be in cases when the distribution of the dependent variable is difficult to model and approximation with standard distributions gives poor results.

Similarly, two main disadvantages of using artificial neural networks in insurance rate making compared to GLM can be identified: time consumption and difficulty of explaining the predictions obtained.

The process of finding a suitable artificial neural network given the available data necessarily involves cross-validation of a range of candidate models and parameter tuning. As seen in this thesis, the more complex the structure of an MLP is, the more time it takes to fit. Depending on model complexity

and computing power, the time required fit an MLP is measured in hours or days, while fitting a GLM takes seconds or minutes. The time consumption of fitting a neural network is also largely dependent on the size of the training data. This implies a restriction in the refinement of the variable grouping among the explanatory variables. With a larger number of variable groups, the aggregated data set increases significantly in size and hence the time consumption of the learning algorithm increases. As for GLM, the time required to fit a model does not change significantly with a larger number of variable groups. Furthermore, in this thesis the final MLP model has a large bootstrapped estimate of variance in test MSE. This indicates that the model needs to be refitted to current data rather frequently, as more information about the policyholders' profiles and claims is obtained. Data updates naturally calls for refitting of GLM models as well. However, this is a much less time consuming project.

The second practical disadvantage of using artificial neural networks in rate making is that it is difficult to explain how a specific risk premium estimate has been obtained. I.e. it is not obvious what role the different explanatory variables play in a fitted artificial neural network. Hence, in the case that policyholders are interested in e.g. why their premium has increased from one year to another, this is a difficult question to answer. Due to the difficulty of explaining the outcomes from an artificial neural network, these are sometimes referred to as "black boxes". This is in direct contrast to the multiplicative factors obtained by fitting a log-link GLM. With a log-link GLM, the effect of the different explanatory variables on a prediction could hardly be more easily understood. In this aspect there is also a difference between a Tweedie GLM and a Poisson-Gamma GLM. The latter gives multiplicative factors of both claim frequencies and claim severities.

5.3 Model improvements and future work

As discussed before, the candidate MLP models selected for cross-validation in this thesis are rather basic in their construction. The limiting factor has been time and computing power. If increasing supply of these resources, tests can be performed on multilayer perceptrons with more sophisticated configurations. Since this thesis shows promising results for using MLP models in car damage insurance rate making, there are good hopes that better performing MLP models in terms of precision and fairness can be obtained. For example, it would be interesting to fit MLP models with a larger number of hidden layers, although a one-hidden layer configuration is motivated by

the universal approximation theorem. It would also be interesting to use different activation functions for different hidden layers.

Also, the selection process of the MLP parameters `tol_level`, `learn_rate` and `batch_size` would benefit from more computing power. Recall that in this thesis, `tol_level` was cross-validated separately from `learn_rate` and `batch_size` due to time limitations. If selecting parameters from a 3-dimensional grid, the number of parameter configurations to cross-validate grows cubically with the number of values for each parameter. Hence, more computer power would enable cross-validating the parameters together, which is to prefer.

To make an as fair as possible comparison between the models, no variable selection methods have been used in this thesis. Before implementing a model in practice, naturally the significance of explanatory variables used must be established. Also, the grouping of explanatory variables affects the model performance and should be further analyzed.

Since GLM has many practical advantages, future work could also include combining artificial neural networks and GLM in the rate making process, by using neural networks as an integrated part of a GLM. It is close at hand to suggest using a neural network for clustering of explanatory variables before fitting a GLM to the aggregated data set. An MLP could also be used to categorize policyholders in different risk segments based on personal information.

6 Conclusion

The results from the final MLP model in this thesis are comparable to the results from the two GLM models. However, the Tweedie GLM performed better on the test set in terms of both fairness and precision. Hence, the Tweedie GLM was more successful in modeling general patterns in the data than the MLP model, despite the latter's ability of modeling high-dimensional non-linearities. As for the Poisson-Gamma GLM, the MLP model showed better precision while the Poisson-Gamma GLM was considered more fair.

Nevertheless, given the restrictions in computing power and time as well as the reasonably basic MLP configurations tested in this thesis, these pioneering tests of using an artificial neural network in car damage insurance

rate making should be seen as successful. With extended computing power and time, an MLP model should have good possibilities of performing in line, or better, than a GLM model.

From a practical perspective, GLM is preferable to artificial neural networks for rate making. This is due to the less extensive process of fitting a model to current data as well as simplicity of understanding how a specific set of explanatory variables gives a certain risk premium.

References

Anderson, D. et al. "A Practitioner's Guide to Generalized Linear Models". In: *2004 Discussion Paper Program - Applying and Evaluating Generalized Linear Models Including Research Papers on the Valuation of P&C Insurance Companies*. Casualty Actuarial Society, 2004, pp. 1–116.

Cybenko, G. "Approximation by Superpositions of a Sigmoidal Function". In: *Mathematics of Control, Signals, and Systems* 2 (1989), pp. 303–314.

Dalkilic, T. E., Tank, F., and Kula, K. S. "Neural networks approach for determining total claim amounts in insurance". In: *Insurance: Mathematics and Economics* 45 (2009), pp. 236–241.

Du, K. L. and Swamy, M. N. S. *Neural Networks and Statistical Learning*. Springer, 2014.

Dugas, C. et al. "Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking". In: *Intelligent and Other Computational Techniques in Insurance: Theory and Applications*. Ed. by Shapiro, A.F. and Jain, L.C. World Scientific Publishing, 2003. Chap. 4, pp. 137–197.

Francis, L. "Neural Networks Demystified". In: *2001 Winter Forum, Ratemaking Discussion Papers and Data Management/Data Quality/Data Technology Call Papers*. Casualty Actuarial Society, 2001, pp. 253–320.

Haigh, J. *Probability Models*. Springer, 2013.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer, 2009.

Mano, C. and Rasa, E. "A Discussion of Modeling Techniques for Personal Lines Pricing". In: Trans 27th ICA, 2012.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*. Chapman and Hall, 1983.

Ohlsson, E. and Johansson, Björn. *Non-life Insurance Pricing with Generalized Linear Models*. Springer, 2010.

Olsson, U. *Generalized Linear Models*. Studentlitteratur, 2002.

Sarle, W. S. "Neural Networks and Statistical Models". In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994*. SAS Users Group International Conference, 1994, pp. 1–13.

Silva, I. N. da et al. *Artificial Neural Networks - A Practical Course*. Springer, 2017.

Zadeh, L. A. "Fuzzy Sets". In: *Information and control* 8 (1965), pp. 338–353.

TRITA -MAT-E 2017:28
ISRN -KTH/MAT/E--17/28--SE