UNIVERSITY OF TWENTE.

Master Industrial Engineering and Management
Financial Engineering and Management

# Exploring Data-Driven Clustering in Generalized Linear Models for Insurance Pricing

Tijmen C. Hommels

M.Sc. Thesis
January 2024

**Supervisors:**
dr. B. Roorda
dr. R.A.M.G. Joosten

Faculty of Behavioural, Management and Social sciences
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

# Acknowledgments

First and foremost, I would like to thank the Dutch insurance company for giving me the opportunity to conduct this research and for the learning opportunities provided. Also, many thanks to the pricing team for making me feel welcome and always helping me with any questions I had. A special thanks to Eelco Zandberg for guiding me thought out this research and being available for advice, feedback and questions.

I am also grateful for Berend Roorda, my supervisor at the University. I would like to thank him for the feedback and advice he provided that allowed me to improve my research. Thanks should also go to Reinoud Joosten for his invaluable feedback on my research. His feedback allowed me to improve the level of this research.

Finally, I would like to thank my family and friends, for all the love and support I received.

# Management summary

It is important for insurances companies to model the risk of their insurance policies accurately. Knowing the risks allows an insurance company to determine individual premium prices that ensure good profitability, competitive market position and risk exposure. The pure premium represents the anticipated cost of the risk that an insurance company faces for a specific policy. These pure premiums are modelled using the industry standard, generalized linear models (GLMs). This is thanks to their ease of implementation and explainability of the outcome. Specific risk factors of the policyholder and the insured object determine the pure premium.

The insurance company currently uses expert-made clusters for each of the individual risk factors to overcome the GLMs shortcomings when facing continuous variables or high cardinal categorical variables. The pricing department at the insurance company is interested in determining whether their current GLM models can be improved by clustering their numerical and categorical risk factors in a data-driven way.

Given that it has been shown that machine learning (ML) models can match and even outperform the standard GLMs in insurance pricing (Henckaerts et al., 2021; Avanzi et al., 2023). The pricing department also wants to investigate these models. Especially if they can be used with the current clustered data that the GLM uses.

The goal of this thesis is to understand the extent to which clustering within variables improves the current modelling method, of using a GLM, for determining the pure premium via the predicted frequency of insurance claims. Furthermore, we look at how the GLM modelling technique compares to machine learning modelling techniques for the pure premium when using clustered data.

This goal was achieved by looking into multiple clustering methods and cluster amounts and their effect on the deviance of the models. This is the evaluation metric used by the insurance company. From these methods, it became apparent that the Fisher-Jenks and K-means clustering methods on the numerical risk factors proved to match the performance of the experts clusters. The categorical variables are clustered using the entity embedding method and can slightly outperform the expert-made judgments. However, this is at the cost of the explainability as these clusters are determined via a black box ML method, namely neural networks.

The new GLM models do not substantially differ in terms of errors and average pure premium when compared to the existing GLM. However, the pure premium of individual policyholders can differ very much. This change can be attributed to a change in coefficients the GLM attributes to certain clusters. However, it is unclear why these large premium changes do not lead to lower errors.

The multiple ML models tested with the clustered inputs all performed worse than the current GLM. This can be seen in unrealistic pure premiums and a larger error rate. However, the tested ML models and the one used in the entity embedding were not optimized via a grid search for this specific data set. This could possibly be one of the reasons why the results were unrealistic.

We recommend that the insurance company keep its current model and clustering techniques. Ultimately, prioritizing the explainability of models within the insurance industry is of more importance than opting for a marginally better model at the cost of the explainability and large shift in the individual pure premiums.

# Contents

# 1 Introduction

## 1.1 Insurance

Insurance is an agreement between an insurance company and a policyholder in which the insurance company provides financial protection against a specified risk. The insurance company will financially compensate the policyholder in the event that losses occur. This is done in exchange for a premium payment from the policyholder. Insurance increases the ability of people and businesses to cope with adverse effects by providing protection against this. Besides risk mitigation, it also offers people peace of mind (Liedtke, 2007).

While insurance is mainly known for providing protection against risk, its role is not limited to this. Ostaszewski (2018) shows that the role of insurance is to also increase rational risk-taking in society. This allows for the pursuit of risky but profitable activities, such as starting a new business.

There are many different risks people can face, and the insurance industry has many types of insurances to cover them. Some examples include life insurance, health insurance and business insurance. While most of the time people can choose whether they want a certain insurance, some countries require certain insurances. In the Netherlands, where this research takes place, for example health insurance is mandatory (Rijksoverheid, 2023). In this research, we focus on a Dutch business-to-business (B2B) insurer.

## 1.2 How insurance works

As mentioned above, a customer pays the insurance company a premium in exchange for financial protection. The amount of premium a policyholder pays the insurance company depends on many different factors. Factors such as the type of risk covered and the policyholder's characteristics. This may lead to unique insurance premiums for different policies, but also between customers in the same policy.

While many factors go into the final premium paid by the customer, we only focus on the pure premium. The pure premium is the premium needed to cover the policyholder's claim risk, it does not cover any other expenses or profits for the insurance company.

Finding the optimal premium is important for insurance companies as it directly affects their profitability, competitive market position, and risk exposure. Charge too much, and customers will leave for a competitor. Charge too little, and the premiums received might not be enough to cover the expected losses of a policy. Accurately predicting the expected losses of a policy gives the insurance company a good indication of what the policy price should be (Henckaerts et al., 2018).

This optimal price should be found for each customer or risk profile, as charging the same price for all policyholders leads to adverse selection. Adverse selection is what happens when the less risky clients leave for a competitor that offers them a lower premium price based on their low-risk profile. This leaves the insurance company with the higher-risk policyholders. The original premium would now be too low to cover the expected losses as the average risk profile of the policy has increased.
To overcome the effects of adverse selection, insurance companies can charge different premiums depending on a customer's risk profile (see Ross, 2022: Denuit and Marechal, 2007). This is done with the help of models.

## 1.3 The project

This research takes place in the pricing department of a Dutch insurance company. This insurance company is called Company X in the rest of this report. Company X offers various financial products and services, including a wide range of insurances, pension products, and asset management. As mentioned above, we focus on B2B insurance, more specifically an insurance policy for a subset of vehicles. Historical data of the previous 10 years on this policy provides us with data on roughly 700.000 policies.

Many insurance companies, including Company X, use historical data to construct pricing models for their policies (see Henckaerts et al., 2018). This historical data set is used in the predictive modelling of the pure premium. Company X and many other insurance companies use generalized linear models (GLMs) for their pricing models. These models have become the industry standard. This is due to their advantages in implementation and explainability of the outcome (see Henckaerts et al., 2018).

Important regulatory bodies for this industry in the Netherlands, the Netherlands Authority for the Financial Markets (Autoriteit Financiële Markten, AFM and De Nederlandsche Bank, DNB), also indicate that explainability is an important aspect of models in the insurance sector (see AFM and DNB, 2019).

## 1.4 The problem statement

The types of risk factor data used in the pricing models for the policy of this thesis can be divided into numerical and categorical risk factors. Numerical risk factors can be either discrete or continuous; examples include the age of the policyholder or the amount insured. Categorical risk factors have a discrete number of possibilities, examples include car brands or sectors of work.

Modelling with categorical factors in a GLM can be rather straightforward if the number of options is limited and if each option is adequately represented. Numerical factors prove to be more difficult for GLMs as they can contain many different options. This can lead to inaccurate estimates and a large amount of risk levels for only a few policyholders. One approach includes clustering these numerical and spatial variables into categorical risk factors with a limited number of options in a data-driven way (Henckaerts et al., 2018).

Currently, Company X clusters its numerical factors in a non-data-driven way, with the use of experts. The use of experts is not limited to numerical factors, Company X also uses experts to group their high-cardinality categorical risk factors. The pricing department at Company X is interested in seeing if their current GLM models can be improved by clustering their numerical and categorical risk factors in a data-driven way.

Recently, a lot of research has also been done into models that can replace GLMs, as GLMs do have shortcomings. Henckaerts et al., 2021 and Avanzi et al., 2023 showed that machine learning (ML) models can match and even outperform the standard GLMs in insurance pricing. These other types of models do have some disadvantages that make insurance companies less likely to adopt them over the standard GLM, an important one is the lower explainability of the outcome.

As mentioned previously, the regulatory bodies of DNB and AFM find it important that insurance companies take explainability into account in their models. For these reasons, Company X still intends to use models based on GLMs. However, in the future, they might want to use a model based on ML. Due to this, Company X wants to know how or if a ML based model using data-driven clusters can improve their prediction of the pure premium.

Thus, the research goal of using these methods is to understand the extent to which clustering within variables improves the current modelling method, of using a GLM, for determining the pure premium. Furthermore, we look at how the GLM modelling technique compares to other more novel and future forward machine learning modelling techniques for the pure premium when using clustered data.
From this follows the research question, described in the section below.

1. Which are the current methods to handle numerical and categorical data for GLMs?

2. How can clusters be made, along with determining their optimal number, to ensure that each cluster significantly influences the pure premium?

3. How do combinations and variations of numerical and categorical clustering impact the pure premium?

4. How does pure premium modelled by the GLM compare to novel modelling techniques?

## 1.5 Methodology

To achieve our goals multiple models are compared in their prediction of the insurance claim frequency of a specific data set, provided by the insurance company. The models contain different combinations of data-driven and expert-made clusters for the chosen risk factors. Multiple different data-driven clustering techniques are tested on both numerical and categorical variables.

These models are combined with a GLM that predicts the severity of claims to see what the effect of the clustering is on the pure premium. The difference in the pure premium between the models are compared. Finally, the industry standard model is compared to different ML models using the same clustered inputs.

## 1.6 Novelty of this research

The novelty of our research is the combination of clustering on different risk factors. Previous research has been done into clustering for numerical and categorical risk factors, however each study has only focused on one of these specified groups. Furthermore, we test some of the well-known clustering techniques on a novel data set provided by a company actively working with it in the insurance industry. Our research could help insurance companies decide on the technique they use to form clusters for their risk factors. The choice of clusters could provide a more accurate model of the pure premium, which could provide customers with a premium price that more accurately reflects their risks.

## 1.7 Organization of the report

Chapter 2 gives background information on insurance, GLMs and different clustering techniques. Chapter 3 outlines the experiments conducted and the design of the different models. We show the results of these experiments in Chapter 4. Chapter 5 discusses the results, and recommendations for future work are also given. The conclusions are given in Chapter 6.

# 2 Background

In this chapter, we give a brief introduction and history on insurance. Then a brief explanation is given on different factors that influence people's decisions about needing insurance and how an insurance price is calculated. The different models used to calculate the price of insurance are then further highlighted in more detail. Finally, different techniques on how to group data together are explained.

## 2.1 History of insurance

Insurance has been around for a long time. The first forms of insurance were practised by Chinese, Indian and other traders in the second and third centuries BC (Trenerry, 1926). Here, the traders distributed their risk of losing their trading goods by distributing them over multiple trading vessels.

For the traders, the risk was losing their trading goods. More generally, risk, describes the chance of an occurrence happening that has an unfavourable outcome. This could be loss or damage to an asset, or even death or injury to a person. The amount of risk someone is exposed to depends on how likely the occurrence with the unfavourable outcome is to happen, also known as the frequency, and how severe the outcome of this occurrence is, known as the severity (Henckaerts et al., 2018). The risk can be described by a combination of frequency and severity. This means that the amount of risk someone is exposed to can vary depending on the situation. For instance, the chance of being bitten by a snake is higher if you live in a country with snakes.

Insurance has evolved over the centuries since the traders in the second and third centuries BC. The more modern form of insurance, where a monetary premium is paid to cover a certain risk, was developed around the 14th century, when maritime insurances were offered to traders. The premiums of these insurance contracts varied with the risk. Over the years, insurance contracts were made to cover more than trading risk. This includes guilds that protected their members against loss from fires and shipwrecks. Other types of insurance also include life insurance, which pays a certain amount of money if the person passes away in a given time frame and was first seen in 1536 (Masci, 2011).

In current times, insurance is offered for all types of risks. This can range from the more traditional risks covered by insurances such as health insurance, property insurances, and car insurance (Kilroy, 2022) to more specialized risks, an example includes insurances on body parts (Kirkpatrick, 2022).

The contract between the insurance company covering the risk and the insured entity is called the insurance policy. The insurance policy specifies which risks are covered and what the insured entity has to pay, also called the premium. Insurance policies can be offered to many different types of policyholders. These could be consumers, business-to-business or even insurances companies themselves in the form of reinsurance. Reinsurance contracts allow the insurance company to transfer a portion of their risk to the reinsurance company (Blazenko, 1986).

The amount of premium a policyholder pays the insurance company depends on many different factors. Factors such as the type of risk covered and the policyholder's characteristics. This leads to unique insurance premiums for different policies, but also between customers in the same policy.

## 2.2 Science of insurance

The development of insurance types over the centuries has gone hand in hand with the development of the psychology behind insurances and the mathematical framework to find the appropriate insurance premium. While we focus on the insurance side of this framework, a lot of research has also been done on the entity's side of determining if insurance is worth it (Kaas et al., 2008; Schlesinger, 2013).

In the following section, we first show why people want insurance and how many different factors play a role in determining whether an entity wants to be insured. Factors include, among others, a person's culture (H. Park et al., 2002) and wealth (Kaas et al., 2008). Following that, the insurance industry's side of how premiums are priced is shown.

### 2.2.1 Why do people want insurance?

People are willing to pay a price for being insured. The mathematical theory behind this is that people attach a certain value to their wealth, instead of their wealth directly. This value can be expressed using a utility function, $u(w)$, where $u()$ is the utility function and $w$ the actual wealth of the person. The theory states that when a person faces a decision, they choose the option with the highest expected utility (Kaas et al., 2008). When someone is faced with a risk, they can calculate their expected utility and how much they would pay to insure this risk by using Equation 1. $X$ is the random loss of the risk, this describes the chance of losing a certain amount of wealth. An example could be (50% no loss of wealth; 20% loss of €100; 30% loss of €500). Equation 1 shows that for a certain range of insurance premiums, people are willing to insure themselves against a risk. Here, $P$ is the premium paid for the insurance.

$$E[u(w - X)] = u(w - P) \tag{1}$$

While it is impossible to determine the utility function of a person, it has some properties. It is assumed that more wealth would lead to a higher utility level, which means $u()$ should be a non-decreasing function. A further assumption about the utility theory is that it should be marginally decreasing (Kaas et al., 2008). This means that additional wealth increases the utility function less when you already have a lot of wealth. Someone earning minimum wage is a lot happier with a €100 bonus than a wealthy man like Jeff Bezos would be.

Aside from mathematical theory, other sociopolitical and cultural factors also significantly influence the popularity of insurance in a nation (H. Park et al., 2002). These influences can be economic, demographic, and institutional variables. S.C. Park shows how cultural aspects influence the consumption of non-life insurance (S. C. Park and Lemaire, 2012). They also show that this consumption of insurance can be affected by the population's religion. This indicates that research done on the insurance pricing of a specific policy could also be region-specific.

### 2.2.2 Pricing of premiums

The premium a policyholder pays is built up from multiple components, split up in two sections: pure premium and other costs. The pure premium is the component that covers the risk of the insurance policy. The other costs cover the expenses made by the insurance company, these can be wages, marketing costs, profit, and others. We focus on the pure premium.

Many different methods are used to determine the pure premium of a policy. Most methods include the use of models with historical data (Henckaerts et al., 2018). In some cases, expert opinion might be the best way to find a pure premium. Here, the expert uses their knowledge and previous experience to find the expected cost of claims per insured entity. This can happen if not a lot of historic data is available on the risk of the policy (Charters de Azevedo et al., 2016).

As mentioned in Section 2.1, the risk depends on the frequency and severity of an occurrence with an unfavourable outcome. Due to the nature of this, the risk can be modelled in two different ways. Either by modelling the risk in one model or by creating two models, one for the frequency and one for the severity. A disadvantage of using models can be the lack of (historical) data. When using one model, the goal of the insurance company is to model the expected cost of claims per insured unit directly, as performed by Jørgensen and Paes De Souza, 1994

When using two models, the insurance companies model the expected number of claims in a given time period (frequency) and the expected size of the claims (severity) separately. Frequency and severity are typically assumed to be independent (see Henckaerts et al., 2018). Due to the assumed independence of the frequency and severity, the two can be multiplied to calculate the pure premium (see E. W. Frees, 2014). This is shown in Equation 2, where $F$ is the frequency, $S$ is the severity, and $PP$ is the pure premium.

$$PP = F * S \tag{2}$$

## 2.3 Models for pure premium

There are many different models that can be used to predict the pure premium. The standard model to calculate the pure premium for insurance companies is the generalized linear model (GLM) (see Haberman and Renshaw, 1996; de Jong and Heller, 2008). The Generalized linear mixed model (GLMM), an extension of the GLM (see Breslow and Clayton, 1993), can also be used (see Avanzi et al., 2023). Recently, it has also been shown that machine learning (ML) models can be used to predict the pure premium. These models can be tree-based such as random forests (see Henckaerts et al., 2021), boosted models such as gradient boosted (GB) (see Yang et al., 2016) or extreme gradient boosting (XGBoost) (see Gao et al., 2018) Another ML model that can be used is a neural network (NN) (see Wüthrich and Merz, 2019).

## 2.4 GLMs / regression models

GLMs have become the industry standard way to model the pure premium. The reason these models are the industry standard is because of their advantages in implementation and explainability of the outcome (see Henckaerts et al., 2018). Regulatory bodies for the insurance industry, such as the Netherlands Authority for the Financial Markets (Autoriteit Financiële Markten, AFM) and De Nederlandsche Bank (DNB), also indicate that explainability is an important aspect of models in the insurance sector (see AFM and DNB, 2019).

### 2.4.1 Linear models

GLMs are a form of regression modelling, a change in one variable is explained by the change in other variables. The classical linear model forms the basis of the GLM (de Jong and Heller, 2008). The linear model describes the relationship between the response variable $y$ and the explanatory variables $x_1, x_2, ...x_p$, where p is the amount of explanatory variables.

The linear relation can be expressed by Equation 3. Here $\beta$ are the regression coefficients, with $\beta_0$ being called the intercept. $\epsilon$ is the disturbance that adds noise to the linear relation between the response and the explanatory variables. This variable represents all other factors that influence the response variable $y$ other than the explanatory variables. The linear model assumes that disturbances $\epsilon$ are independent and normally distributed: $\epsilon \sim N(0, \sigma^2)$. The linear model does not assume that the response variable $y$ is normally distributed (de Jong and Heller, 2008).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon \tag{3}$$

Equation 3 can also be written in matrix notation. $\boldsymbol{y}$ is an 1 x $n$ vector of response variables, where n is the number of data points. $\boldsymbol{X}$ is a ($n$ x $p$) matrix containing the explanatory variables for n data points. $\boldsymbol{\beta}$ is a (1 x $(p+1)$) vector containing the regression coefficients. $\boldsymbol{\epsilon}$ is a (1 x $n$) vector of disturbances. Fitting a linear model requires estimating the regression coefficients, $\boldsymbol{\beta}$, such that the disturbances, $\boldsymbol{\epsilon}$, are minimized. This can be done using the least squared method, in which the sum of squared errors, $\|\boldsymbol{\epsilon}\|_2^2$, is used as a target to minimize.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4}$$

### 2.4.2 GLM models

While the linear model does allow for the modelling of the relationship between the response variables and the explanatory variables, it is limited by the assumption that the disturbances are normally distributed. For generalized linear models, this assumption is relaxed. This allows for response variables that have distribution models other than a normal distribution (de Jong and Heller, 2008).

With insurance data, the assumptions of the normal model are frequently not applicable, thus GLMs are preferred over the linear model. Frequency of insurance claims is assumed to be Poisson distributed and the severity is assumed to be gamma distributed (Coutts, 1985).

The GLM expands on the linear model by assuming that the response variable is generated from a distribution in the exponential family. This allows the response variable to have a binomial, Poisson, or gamma distribution. The relationship between the mean of the response variable, $\mu$, is described in Equation 5. Here g is the link function, this determines how the mean of the response relates to the explanatory variables $x$. $(X\beta)$ is known as the linear predictor and describes the relationship between the regression coefficients and the explanatory variables: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$, as seen in the classical linear model (de Jong and Heller, 2008).

$$\boldsymbol{E(Y|X) = \mu = g^{-1}(X\beta)} \tag{5}$$

The linking function needs to be chosen such that Equation 5 holds. This depends on the distribution of the response variables. As stated earlier, the distribution of interest in the insurance pricing sector are Poisson and gamma distributions. The linking function and mean functions for these are shown in Equations 6 and 7 respectively (de Jong and Heller, 2008).

$$g(\mu) = ln(\mu) = X\beta \rightarrow \mu = exp(X\beta) \tag{6}$$

$$g(\mu) = -(\mu)^{-1} = X\beta \rightarrow \mu = -(X\beta)^{-1} \tag{7}$$

Using this information, a GLM can be constructed to find the estimated value of the response variable. In Equation 11, a GLM is shown that assumes the response variable has a Poisson distribution, which is a distribution of interest in the insurance pricing sector. Due to the choice in the link function, the effects of the explanatory variables and their regression coefficients are multiplicative instead of the additive effects seen in the classical linear model and shown in Equation 3. Fitting is used to find the regression coefficients, $\boldsymbol{\beta}$. This is done via maximum likelihood estimations. One way to do this is by iteratively solving to minimize a loss function. For GLMs, the minimization problem is shown in Equation 8. Here n is the amount of samples and $\alpha$ is the L2 regularization penalty. $||\beta||_2^2$ is the square of the magnitude of the regression coefficients $\beta$. $d$ is the deviance [1]. For a Poisson log link function the deviance is given by Equation 9, and for the gamma log link function the deviance can be found using Equation 10. Here, $y_i$ is the target and $\hat{y_i}$ is the estimated value from the models (de Jong and Heller, 2008).

$$min_\beta \frac{1}{2n} \sum_i^n d(y_i, \hat{y_i}) + \frac{\alpha}{2} ||\beta||_2^2 \tag{8}$$

$$d(y_i, \hat{y_i})_{Poisson} = 2(y * log(\frac{y}{\hat{y}}) - y + \hat{y}) \tag{9}$$

$$d(y_i, \hat{y_i})_{Gamma} = 2(log(\frac{y}{\hat{y}}) + \frac{y}{\hat{y}} - 1) \tag{10}$$

$$\begin{aligned} \mu &= exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \\ &= exp(\beta_0) * exp(\beta_1 x_1) * exp(\beta_2 x_2) * ... * exp(\beta_p x_p) \end{aligned} \tag{11}$$

The strength of the GLM model is that it is flexible in the amount of features it receives. The individual effect of each feature is also easily traceable, making the results easy to understand. Furthermore, GLMs can handle different types of data, including continuous data and categorical data, if this data is one-hot encoded (explained in Section 2.6.2).

### 2.4.3 Disadvantages of GLMs

While GLMs allow for the modelling of a relationship between explanatory variables and the response variable, they do have some shortcomings. Among them is the assumption that risk factors relate to the response (the calculated pure premium) in a linear way. This might not always be the case for continuous risk factors, which means the GLMs may not accurately capture the nonlinear relationship, thus leading to inaccurate predictions. Other shortcomings include the difficulty of modelling continuous and spatial variables (see Henckaerts et al., 2018).

[1] The term deviance can have different definitions based on the context in which it is used. We use the definition used in statistics. Deviance: A measure for judging the extent to which a model explains the variation in a set of data when the parameter estimation is carried out using the method of maximum likelihood (Upton and Cook, 2008).

While categorical factors with a low number of options can be handled quite well by a GLM, high-cardinality categorical features do offer some issues, such as sparse data at certain levels, which leads to more uncertainty in parameter estimates and predictions (see Avanzi et al., 2023). Another issue is that the large number of levels leads to a higher computational resource requirement, as indicated by Shi and Shi, 2022. One approach looked into in different research papers includes clustering these high-cardinality categorical features into categorical risk factors with a limited number of options (see Henckaerts et al., 2018; E. Frees and Valdez, 2008).

## 2.5 ML models

As mentioned in Section 2.3, different machine learning models can also be used in the prediction of the pure premium. These types of models can offer better results than the standard GLM, as shown by Henckaerts et al., 2021, Wüthrich and Merz, 2019 and Wuthrich and Buser, 2016. However, ML models do have some disadvantages that make insurance companies less likely to adopt them over the standard GLM. An important disadvantage is the lower explainability of the outcome. As mentioned previously, the regulatory bodies of DNB and AFM find it important that insurance companies take this explainability into account. In the following sections, different ML models are highlighted.

### 2.5.1 Tree based model

The most basic tree based model is a decision tree first proposed by Quinlan, 1986. Decision trees try to predict a target by following a set of rules it has learned from the set of data features. These sets of rules split the input space, all inputs and their possible features, into sub spaces. Each of these sub spaces is assigned a constant prediction value. An example of a decision tree is shown in Figure 1. The decision tree contains a root node, branches, internal nodes and leaves. The root node is the first decision of the tree, in Figure 1 this is the "Is there swell?" node. This node branches out into an internal node "wind" and a leaf. A leaf indicates the end of a branch and contains a constant prediction, the leaf found by answering "no" to the root node tells you not to go surfing (IBM, 2023).

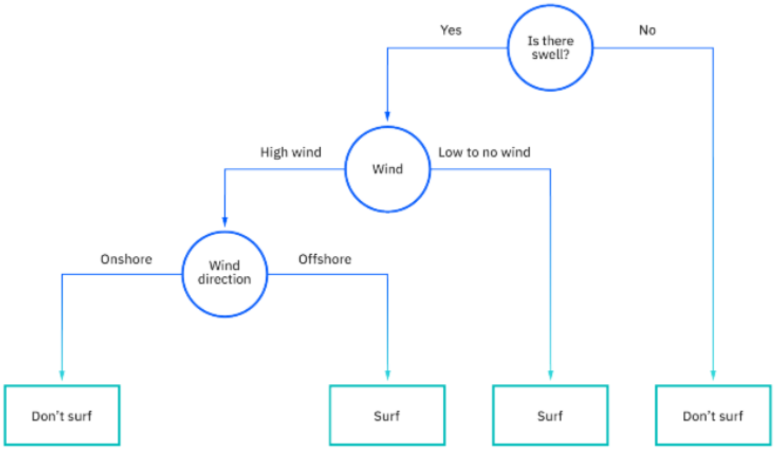


Figure 1: Decision tree on if you should go surfing (IBM, 2023)

Finding suitable decision nodes is done by using the impurity criterion. One of the most common methods is using entropy for the impurity criteria (Louppe, 2015). Equation 12 shows how entropy is defined. S is the data set over which entropy is calculated, C represents the classes in data set S, and $p(c)$ is the portion of data points that belong to class c in the data set S. When a data set contains only 1 class, the entropy is 0. For regression models, the impurity function is based on the squared error loss shown in Equation 13. Here $y_i$ is the target of data point i, $\hat{y}$ is the mean value of the node, and $n_s$ is the number of data points in the node (Louppe, 2015).

$$Entropy(S) = -\sum_{c \in C} p(c) log_2 p(c) \tag{12}$$

$$Impurity(S) = \frac{1}{n_s} \sum_{i}^{n_s} (y_i - \hat{y})^2 \tag{13}$$

Entropy can be used to find the information gain of a node, as it represents the difference before and after a split. The information gain can be found using Equation 14. Here $IG$ is the information gain, $a$ is a specific attribute, $Values(a)$ is all the possible values of attribute $a$, and $Entropy(S)$ is the entropy of the data set present in the node before splitting. The fraction, $\frac{|S_v|}{|S|}$, represents the portion of the data set S that has gone into the new node or leaf, and $Entropy(S_v)$ is the entropy of that node or leaf. The best split is the one with the highest information gain (IBM, 2023). The method of making a decision tree by making nodes that locally, without regard of future nodes, gain the most information is called the greedy method (Louppe, 2015).

$$IG(S, a) = Entropy(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{14}$$

Some advantages of decision trees are that they are simple to understand and interpret, this is enhanced by the possibilities of visualizing the trees. Furthermore, it can handle numerical and categorical data with little to no preparation, such as normalization. Decision trees also have clear disadvantages. This includes the problem that they can create very complex trees, leading to overfitting the data. Another issue is that decision trees are unstable, as small variations in the data can result in a completely different tree (ScikitLearn, 2023b).

To combat the issues of the high variance of decision trees, Breiman proposed the Random forest method (Breiman, 2001). Random forest perform random sampling with replacements to create new data set, this is called bootstrapping (Efron and Tibshirani, 1994). For each new data set a decision tree is made, however at each node, features are randomly picked that are considered for the best split at that node (Guo and Berkhahn, 2016). This also means that not every tree uses the same data or the same features.

The prediction on an unseen sample is the average outcome of all the trained decision trees. Due to averaging over multiple decision trees, the variance is reduced when compared to single decision trees.

### 2.5.2 Gradient boosted trees

Gradient-boosted trees are an ensemble ML method (see Friedman, 2001). These types of models combine multiple models to create one predictive model. Gradient boosted trees combines multiple decision trees, just like the random forest.

Gradient tree boosting uses a set amount of trees. Each tree tries to improve on the errors made by the previous ones. Each iteration, a new decision tree is trained on the pseudo-residuals of the previous trees. The pseudo-residuals, $\gamma$, for a given loss function, $L$, are shown in Equation 15. $f(x_i)$ is the prediction of the model based on the previous m-1 trees. When using a squared error loss function, the m-th decision tree fits on the residual of $y_1 - f_{m-1}(x_i)$ (Friedman, 2001).

$$\gamma_{im} = -[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)}]_{f=f_{m-1}} \tag{15}$$

The advantage of gradient boosted trees is that they are less prone to over fitting when compared to using just one decision tree. Furthermore, due to the process of trying to correct all previous trees it also minimizes the bias of the model and not only the variance, like in a random forest model (Guo and Berkhahn, 2016). Boosted trees are however less transparent than GLMs (Avanzi et al., 2023).

A popular implementation of gradient-boosted trees is eXtreme Gradient Boosting (XGBoost), first introduced by Chen and Guestrin, 2016. XGBoost incorporates a regularized model to prevent overfitting and to control model complexity. Furthermore, it is designed to be an efficient model that can handle large data sets using fewer computing resources compared to other tree boosting systems (Chen and Guestrin, 2016).

### 2.5.3 NN models

NN are models that have been inspired by how the brain works, specifically the neuron system. NN has been successfully used in many different applications and is the basis for many complex ML models (Schmidhuber, 2015; Fujimoto et al., 2018).

NN consists of multiple layers, with each layer containing a certain amount of neurons. NN contains an input layer, an output layer and one or more hidden layers, and the number of neurons can vary per layer. Figure 2.5.3 shows a NN with three inputs, a hidden layer with four neurons and two outputs.

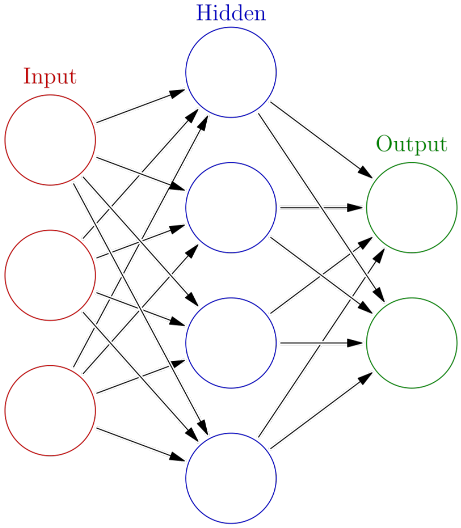


Figure 2: NN with three layers, each containing multiple nodes (Wikipedia, 2023)

Each node can be seen as its own linear regression model, as each node contains inputs, weights and biases, the only exception is the node in the input layer. Equation 16 shows how the output of each neuron is computed. Here, $w$ indicates the weights that are applied to the inputs $x$, $w_0$ is called the bias. The weights help indicate the importance of a given input. $x$ are the inputs received from other neurons. The amount of inputs, N, a neuron receives depends on how many neurons it is connected to.

$$y = \sigma(w_0 + w_1x_1 + w_2x_2 + ... + w_Nx_N) \tag{16}$$

$\sigma$ is the activation function of the neuron, this is the same for all neurons in a layer. The neuron applies the activation function over the sum of the weighted input to decide what its output is. A common choice for the activation function is the Rectified Linear Unit (ReLU) function (Guo and Berkhahn, 2016). This is a nonlinear function that is equal to zero for inputs under a desired threshold (usually 0) and linear with the input if it is bigger than the threshold. These 0 outputs help the NN reach its best performance (Glorot et al., 2010).

To train the NN a loss function is needed. For regression models, this is the mean squared error between the outcome of the NN and the observed value of the training data. The goal of training is to minimize the loss function by adjusting the weights and biases of each neuron. The process of updating weights is done via the stochastic gradient descent method (Robbins, 1951). This method iteratively updates the weights and biases based on the loss function.

Just like with the GLM, NNs can handle many different types, but need some preprocessing when using categorical data. One of the methods NNs can use to handle categorical data is entity embedding. Entity embedding allows the NN to learn relationships between categories, this is explained further in Section 2.6.2 (Guo and Berkhahn, 2016). The advantage a NN has, is that it can find non-linear relationships in data and perform complex pattern recognition. This makes it suitable for many different tasks, including claim predictions for insurances. The ability to do these complex tasks does come with downsides. NNs are complex and need a lot of data for their training. Their complexity also makes NNs difficult to interpret as to why they make specific predictions (Kuhn and Johnson, 2013).

## 2.6 Clustering

Clustering is a method of grouping objects together in such a way that objects in each cluster are similar. Clustering has been applied to many different fields, from engineering and science to everyday life (Anderberg, 2014). In insurance pricing, it can be useful to place similar variables into bins to reduce the complexity of certain models when faced with high-cardinality variables (Henckaerts et al., 2018). Furthermore, it can help to place continuous variables into certain clusters to ensure better fitting over the whole range of the variables.

Different types of data use different types of clustering techniques. The types of data used in the models for this thesis can be divided into numerical and categorical factors. Numerical risk factors can be either discrete or continuous; examples include the age of the policyholder or the amount insured. Categorical risk factors have a discrete number of possibilities; examples include car brands or sectors of work.

### 2.6.1 Numerical data

For numerical clustering of a one-dimensional array, different techniques exist. Some of the most popular techniques are highlighted below (Henckaerts et al., 2018).

- **Uniform**: This method is also called equal-width binning. In this method, $k$ amount of bins are created with identical widths. The width of these bins is found by using the minimum and maximum values of the observations, $o$, for all $n$ observations. The width of each bin is: $\frac{max(o)-min(o)}{k}$. This method works well for uniformly distributed data. It does, however, perform poorly for skewed data (Henckaerts et al., 2018).

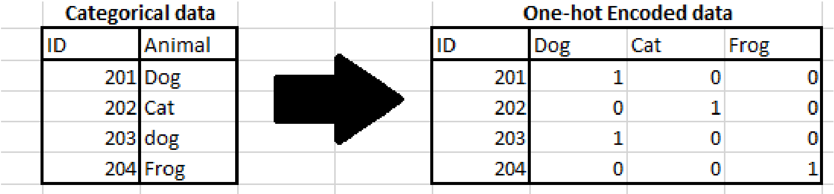
- **Quantile**: This method tries to ensure that each bin has the same number of observations. The expected amount of observations in each bin is approximately $\frac{n}{k}$, where $n$ is the total amount of observations and $k$ is the desired amount of bins.

- **Fisher-Jenks algorithm**: This is a method introduced by Jenks (Jenks, 1967). This data clustering method is designed to determine the best clustering of values into a given amount of clusters such that the variance within groups is minimized and maximizing the distance between the means of other clusters (Fisher, 1958). Using an iterative process, bins are created such that the sum of squared distances between observations and the mean of the respective bins is minimized for all bins.

- **K-means**: This is an iterative clustering method developed by Lloyd (Lloyd, 1982). This method updates the centre location of the k cluster by taking the mean of the observations assigned to the cluster. From here, the observations are reassigned to the cluster with the closest centre location. This process is repeated until the centres of the clusters do not move significantly (MacQueen, 1967).

### 2.6.2 Categorical data

The standard way to use categorical data in GLMs and NNs is the use of one-hot encoding (also known as using dummy variables) (Avanzi et al., 2023; Rodríguez et al., 2018; Shi and Shi, 2022). This technique transforms categorical data into binary data. One-hot encoding creates a new variable for each unique feature of a categorical variable. Only the new variables corresponding to the observed category are filled in with a one, the rest of the new variables are filled in with zeros (Rodríguez et al., 2018). Figure 3 illustrates how one-hot encoding works.



**Categorical data**

| ID | Animal |
|---|---|
| 201 | Dog |
| 202 | Cat |
| 203 | dog |
| 204 | Frog |

**One-hot Encoded data**

| ID | Dog | Cat | Frog |
|---|---|---|---|
| 201 | 1 | 0 | 0 |
| 202 | 0 | 1 | 0 |
| 203 | 1 | 0 | 0 |
| 204 | 0 | 0 | 1 |

Figure 3: Example of one-hot encoding categorical data

Categorical variables with a lot of features, also known as high-cardinality categorical features, generate a lot of new variables when using one-hot encoding. GLM and NNs face some issues with one-hot encoded high-cardinality categorical features. These include computational burden and estimation uncertainty, as some features do not have a lot of data (Shi and Shi, 2022; Avanzi et al., 2023).

Multiple different approaches have been tried in the modelling of high-cardinality categorical features, such as grouping the levels with similar risk behaviors; this has been done manually (via experts) and in a data-driven way (see Micci-Barreca, 2001). Other methods include entity embedding from neural networks (see Shi and Shi, 2022).

In Section 2.5 entity embedding was listed as an advantage of the NN model. This is due to the fact that entity embedding is trained using a NN.

The entity embedding method is used to automatically learn the representation of categorical features in multidimensional spaces. Categorical features with similar effects on the model are located close to each other in these multidimensional spaces. An advantage of entity embedding is that the embedded layers can be used as inputs in other models. Furthermore, it can be used for visualizing categorical data and for data clustering (Guo and Berkhahn, 2016).

Entity embedding maps each state of categorical variables into a multidimensional space. This mapping can be seen as an extra layer of linear neurons on top of the one-hot encoded categorical inputs. The dimension of the embedding is a parameter that can be selected. This parameter is bound between 1 and $q_i$ - 1, $q_i$ is the number of unique features for the categorical variables $x_i$. The mappings are the weights of the extra layer of neurons and can thus be trained in the same way other neurons of the NN are trained. This also means that the training of the NN changes the location of each of the unique features of a category in the multidimensional space (Guo and Berkhahn, 2016).

# 3 Methodology

In the background, we have shown the current methods to handle numerical and categorical data for GLMs, thereby answering the first research question.

In this research, we use entity embedding for categorical variables to reduce the reliance on experts. This also contributes to future proofing the pure premium determination while still keeping clarity as necessary by the AFM, DNB, and Company X. For the numerical risk factors we use traditional data-driven clustering techniques. These split techniques of clustering are used to answer Research Question 3, by combining and varying the clusters which serve as input for the GLM. All the clustering techniques are only applied to the frequency data set and consequently the frequency models.

Furthermore, we use the numerical clusters together with the entity embedded clusters to train machine learning models. We compare three machine learning models (NN, XGBoost and RF), these determine the claim frequency, which is essential in finding the pure premium. This is done to answer Research Question 4.

Before we get to explaining the methodology of the three remaining research questions, we explain the data cleaning and exploration steps undertaken in the next section. All the steps explained in the rest of this chapter have been performed in Python (Van Rossum and Drake, 2009).

## 3.1 Data preparation

As mentioned in the introduction, this research uses historical data on a B2B insurance policy, more specifically an insurance policy for a subset of vehicles. This historic data is provided by Company X and contains information on all policies of the previous 10 years, this is roughly 700.000 policies. If a policyholder claimed damages, a copy of the policy with the claim information is present in the data set. Each of these policies contains over 100 data fields. These fields range from describing the risk factors of the policy to internal data of Company X on the policy, an example would be the original premium.

The data is first prepared by removing policies that claimed damages but whose claims were less than 10 euros (a threshold set by Company X). As Company X uses the frequency-severity model, two separate data sets are made. The severity data set contained all the remaining damage claim policies. For the frequency data set all copies of a policy, due to claim information, were removed. For each removed copy, the number of claims on the original policy is increased by one.

Most of the data fields in the data sets are not needed in the modelling of the pure premium. For this reason, we selected only 22 data fields for this research. The remaining data fields contained all risk factors used in the modelling of the pure premium, the existing clusters of these factors (made by an expert), data used for identification of the policy and target data. The target data for the frequency set is the frequency of claims, this is found using Equation 17. Here $N_{claims}$ is the amount of claims for a policy and $N_{dur}$ is the number of insurance years for the given policy. The amount of insurance years is always between 0 and 1. The target for the severity data set is how much the payout was for the claim.

$$T_{freq} = \frac{N_{claims}}{N_{dur}} \tag{17}$$

In Table 1 an example of the data is given. This example illustrates how the number of insurance years is always between 0 and 1. As the example shows a policy with an original start date of first of May, every first of May the number of damage free years (DFY) goes up by 1, if no damages have been claimed. As the factors of the policy change, a new row is added tot the data, indicated by Row 2. Another way the factors of a policy change, is by a change in years. This can be seen by comparing Rows 2 and 3. The example shows that rows of policy always change for a new year and at the original start date of a policy.

Both data sets are split into a training set and test set. This split is 75-25 % respectively. All clustering and model training was done on the training data set.

| Row | Policy ID | RF | DFY | Year | End Date | Insurance years | Claims | Target |
|---|---|---|---|---|---|---|---|---|
| 1 | 123456 | ... | 8 | 2018 | 1 May | 0.33 | 0 | 0 |
| 2 | 123456 | ... | 9 | 2018 | 31 Dec | 0.67 | 0 | 0 |
| 3 | 123456 | ... | 9 | 2019 | 1 May | 0.33 | 1 | 3 |

Table 1: Example of the data.

## 3.2 Clustering

In the beginning of this chapter, we mentioned that we focus on the claim frequency modelling. This means that the clustering techniques were all applied to the frequency data set and not the severity data set. Company X uses 9 risk factors in their claim frequency model. 4 of these factors are numerical, while the remaining 5 are categorical risk factors. Out of the 5 categorical risk factors, 2 can be seen as high cardinal with over 75 features present in the risk factor.

### 3.2.1 Numerical clustering

In Section 2.6.1, different numerical clustering techniques were highlighted. The four techniques: Uniform, Quantile, Fisher-Jenks algorithm and K-means were all applied to the 4 numeric risk factors. This was done for cluster sizes ranging from 2 to 34. While the risk factors are numeric, some data points were already assigned to a cluster. These pre-assigned clusters indicated an unknown value. Policies with the pre-assigned value in a specific risk factor were excluded in the clustering of that specific risk factor, this means they kept their pre-assigned cluster.

To evaluate the clusters, the fraction of the variance explained was used. Equation 18 shows how the variance within clusters can be found, known as the within cluster sum squared, $WCSS$. Here $K$ is the number of clusters, $nk$ is the number of data points in cluster $k$, $x_{ik}$ is the value of data point $i$ in cluster $k$ and $\bar{x_k}$ is the mean of the data points in cluster $k$. To find the variance of the total risk factor, $k$ is set to 1. The fraction of the variance explained is then found with Equation 18, this is a value between 0 and 1. Here, $Var_T$ is the variance of the total risk factor.

$$WCSS = \sum_{k}^{K} \sum_{i}^{nk} (x_{ik} - \bar{x_k})^2 \tag{18}$$

$$GOF = \frac{Var_T - WCSS_K}{Var_T} \tag{19}$$

Using the fraction of the variance explained, an optimal combination of clustering technique and cluster amount can be found. This is done using the elbow method (Thorndike, 1953).

### 3.2.2 Categorical clustering

The clustering of the categorical variables was only applied to the two high cardinal variables. The chosen clustering technique for this research is entity embedding. Section 2.6.2 describes how entity embedding works. This method was selected as it captures relationships and similarities between features within the risk factor. This method can also be exported from the NN and used in a GLM.

The NN used to train the embedding of the categorical features is described in Table 2. The size of the embedded space for the features was determined using the following formula $min(\sqrt{n_{feat}}, 50)$. The NN was trained on the frequency targets of the policy, with each policy getting a sample weight equal to the amount of insurance years of that policy. This sample weight also solves the issue of running into infinite frequencies due to dividing by 0 in Equation 17. The other risk factors, in the insurance companies clusters, were also used as input after being transformed using one-hot encoding. This method was used with a NN with a hidden layer size of 120. Another training method was using the continuous risk factors directly, this was done using hidden layer sizes of 10 and 120. This means three different NN training session were performed.

The location of the features of the risk factors in the embedded space was extracted from NN. Based on the average location of these features in 10 embedding spaces, clusters of the data were formed using the K-means method. This research clustered both methods into groups ranging from 2 to 34.

| | |
|---|---|
| Hidden Layers | 2 |
| Hidden Layers Size | 120 or 10 |
| Activation | Relu |
| Output activation | Linear |
| Optimization | Adam |
| Loss | Mean squared error |

Table 2: Variables used in the NN.

## 3.3 Impact of clusters on the GLM

To find the impact of the clusters on the frequency GLM, a base frequency and severity GLM were made. These are GLMs that use Company X's clusters for the four numeric risk factors, and the only cluster made by Company X on one of the high cardinal categorical risk factors. All risk features are then one-hot encoded and used as inputs for the GLMs. Each GLM has its own respective target, and the frequency GLM also uses the amount of insurance years of that policy as weight. The frequency GLM uses a Poisson log link function and the severity GLM uses a Gamma log link function, this is explained in Section 2.4.

In this research, we program the GLMs in Python (Van Rossum and Drake, 2009) using Scikit-learns (ScikitLearn, 2023b) Poisson and Gamma regressor. The Solver is chosen to be 'newton-cholesky' with an $\alpha$, found in Equation 8, of 0.0001 (ScikitLearn, 2023a).

The base GLM for the frequency was then modified by changing Company X's clusters with clusters found by the different clustering techniques. This was done for each numerical risk cluster of Company X with each cluster found for that specific risk factor. The other factors all remained the same as the base GLM. For this, 512 GLMs were made (4 risk factors * 4 methods * 32 clusters). This same method was also applied to the cluster made for the categorical variables. All the GLMs mean Poisson deviance is computed on the test set. This was computed by finding the sum of the deviance of each sample using Equation 9 and then dividing the sum by the total number of data points in the test set.

## 3.4 Impact of combination of clusters on the pure premium

The previous section describes only changing out one existing risk factor for a newly found cluster. Building on that, multiple risk factors were replaced in the GLM. The first combination was made by replacing all numeric risk factors by the optimal risk factors found using the elbow method on the fraction of the variance explained by the clusters.

This GLM was further developed by replacing the two categorical risk factors with the clusters found using entity embedding. We chose to replace the categorical variables with the same amount of clusters as the insurance company had for CV 2, 27 clusters. Finally, a hand-picked combination of the best clusters based on the individual deviance plots was also trained on the GLM. This was done for each of the three methods of training the NN for the clustering of CV 1 and CV 2.

All the trained GLMs were tested on the frequency test set. The GLMs were then compared based on the difference in the following errors: mean absolute error (MAE), mean squared error (MSE), the mean Poisson deviance and the $D^2$. The MAE and MSE are computed using Equation 20 and 21, respectively. The $D^2$ computes the fraction of deviance explained, it can be computed using Equation 22. Here *dev* is Equation 9, $y_{null}$ the optimal prediction of an intercept-only model. A larger $D^2$

indicates a better model, with the maximum possible value being 1.

$$MAE(y, \hat{y}) = abs(y - \hat{y}) \tag{20}$$

$$MSE(y, \hat{y}) = (y - \hat{y})^2 \tag{21}$$

$$D^2(y, \hat{y}) = 1 - \frac{dev(y, \hat{y})}{dev(y, y_{null})} \tag{22}$$

The insurance company uses the Poisson deviance as one of their criteria to assess their GLM models. One of the other criteria is looking into the significance of the risk factors. The selected risk factors for this thesis were all deemed significant by the insurance company. The criteria used by the insurance company to assess their models is not limited to this. An example of one of the many criteria used is that they evaluate whether the coefficients for a risk factor make sense.

For a clearer view on the differences between these GLMs, the frequency results of the test data set were multiplied with the results of the base severity GLM on this same test set. This shows the effect of clustering on the pure premium. The models are compared on the mean pure premium and standard deviation of the pure premium. The minimum and maximum pure premium are also shown.

## 3.5 Comparing the base model with the hand-picked model

We compare the base model and the hand-picked model. This is done by first looking at the weights given for each of the risk factors. Furthermore, the pure premium found using the base GLM model and that of the hand-picked model are compared. This is done by looking at how much the individual premiums change on an absolute and percentage basis. Finally, the difference in average weight for the risk factors of the premiums with the biggest percentage increase and decrease is found.

## 3.6 ML models on GLM

In the background chapter, many different models are shown that can be used to model the claim frequency. We compared different ML models with the base GLM when receiving clustered data. The chosen ML models are RF, XGBoost and NN. The input of these models, the risk factors, were all one-hot encoded. The results of these models were compared using the mean squared error and mean average error. The specifications of the XGB and RF models are shown in Table 3. The NN used the same specification as found in Table 2 with the hidden layer size of 120.

| | XGB | RF |
|---|---|---|
| max depth | 8 | 8 |
| n_estimators | 200 | 200 |
| min child weight | 20 | 1 |
| gamma | 0.01 | 0 |

Table 3: The chosen specifications of the forest ML models.

# 4 Results

We show the results of the different steps undertaken as described in Chapter 3. First, the data preparation results. This is followed by the results of the numerical and categorical clustering. These clusters are placed into the GLM, leading to multiple deviance plots. Afterwards, multiple different GLMs are compared on various errors and a more in depth look at the difference between two of these GLMs is presented. Finally, the results of the different ML models using the original insurance company clusters are shown.

## 4.1 Data preparation

The data preparation resulted in 713570 policies. As mentioned before, the number of damages present in the data set is roughly 1%. The number of damages claimed in each policy is shown in Figure 4. No policy claimed more than 6 damage claims in its insurance years.

Splitting the data into a train and test set results in sets with 535177 and 178393 policies, respectively.

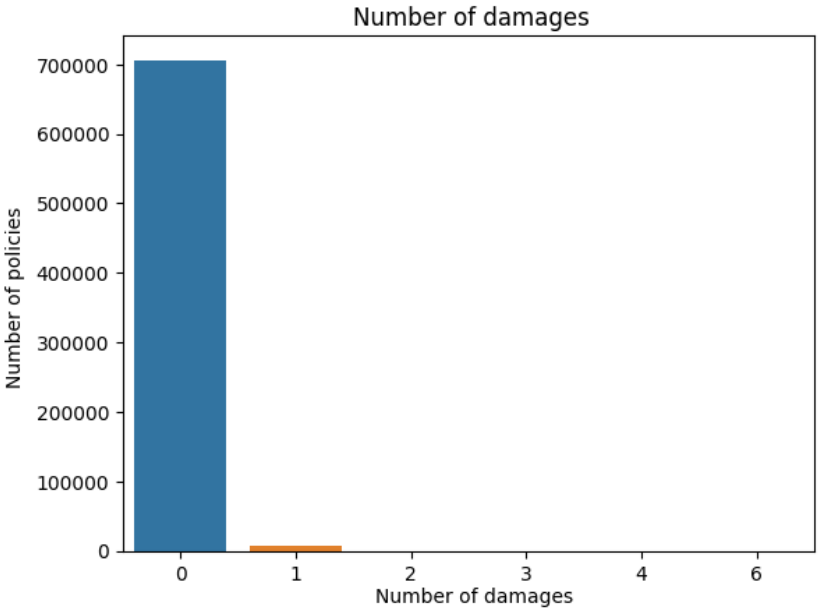


Figure 4: Number of damage claims for each of the policies in the data set

## 4.2 Results Numerical clustering

This section presents the results of the numerical clustering as described in Chapter 3. The results of the clustering are shown for each of the four numeric risk factors (NV 1, NV 2, NV 3 and NV 4). In each of the figures, the goodness of fit for each numerical variable is shown, which was found using Equation 19. The X axis of the graphs shows the amount of clusters made (excluding the pre-assigned clusters). The four different lines in each plot indicate the method used to make the clusters. What method each line represents is seen in the legend of the plot.

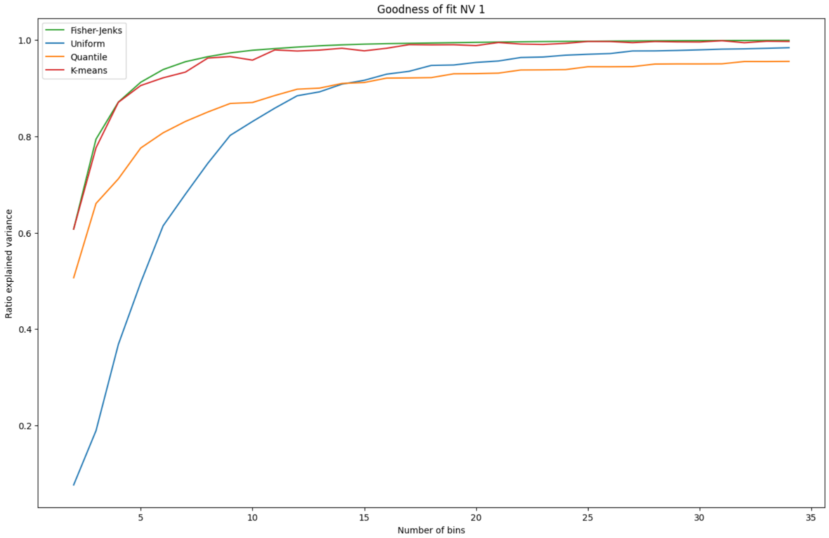


Figure 5: Goodness of fit NV 1.

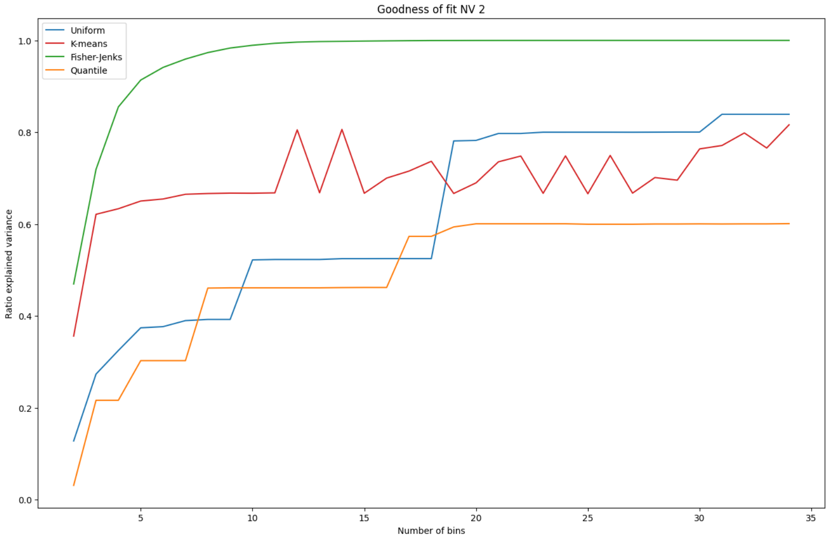


Figure 6: Goodness of fit NV 2.

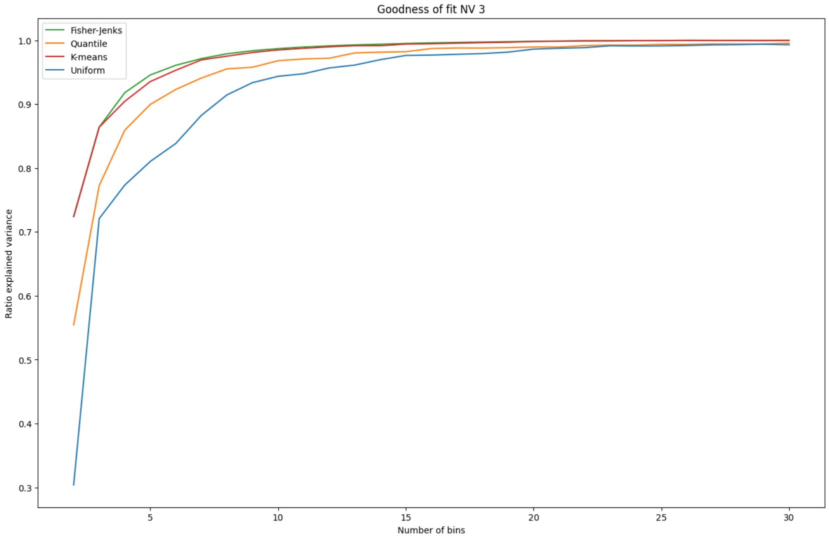


Figure 7: Goodness of fit NV 3.

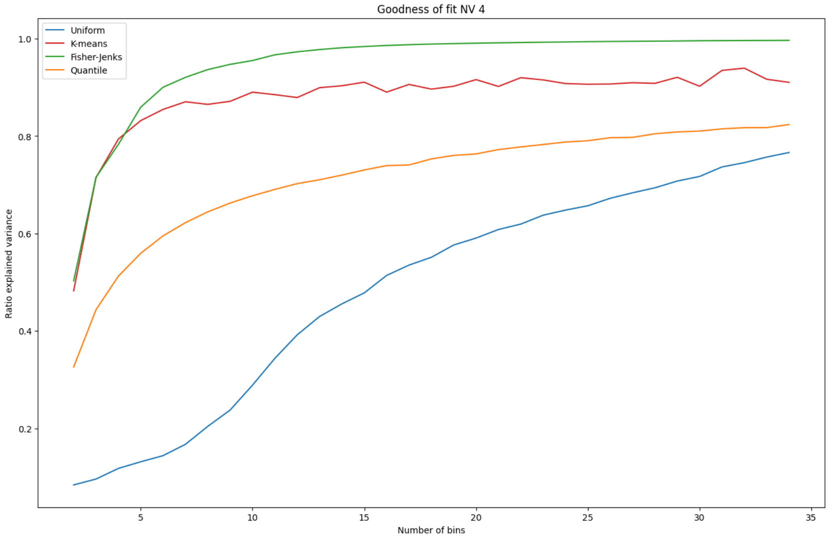


Figure 8: Goodness of fit NV 4.

## 4.3 Categorical clustering

The chosen cluster method for the two high cardinal categorical variables was entity embedding. As mentioned in Chapter 3, entity embedding uses a neural network, the chosen design for this network is shown in Table 2. The two high cardinal categorical variables are known as CV 1 and CV 2. The remaining three categorical variables present are known as IV 1, 2 and 3.

Chapter 3 also shows the different training methods used for the entity embedding. This leads to three different clustering results using entity embedding. The results of the NN with a hidden layer size of 10, trained using continuous inputs, is shown in the figures below.

Results of the NN with hidden layer size of 120 trained on categorical inputs and the results of the NN with the hidden layer size of 120 trained on the continuous variables are shown in Appendix 7.1.

The figures below show the average location of each feature in the embedded space. To visualize the higher order embedded space, the embedding space is reduced to a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008). This average location was obtained from 10 trained NNs. Furthermore, the colour of each plotted feature in the figure represents the cluster that feature belongs to.

### 4.3.1 Categorical variable 1

Figure 9 shows the two-dimension embedded location of the roughly 75 features present in CV 1. This figure shows the results of the clustering of the features into 27 clusters. Each colour of a feature in the figure represents its cluster.

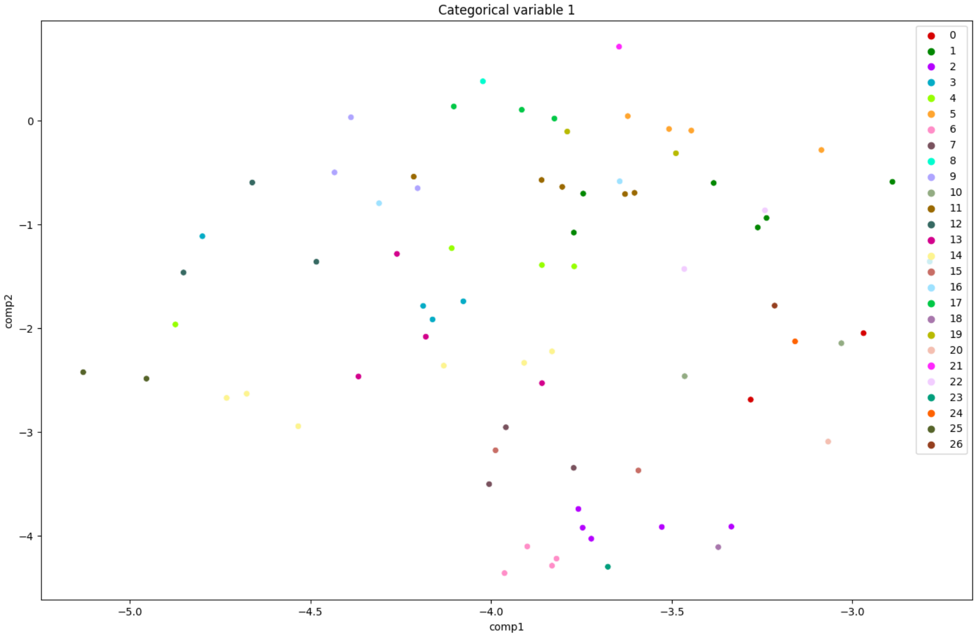


Figure 9: Location of variables in CV 1 in a reduced embedded space. Clustered using the K-means method.

### 4.3.2 Categorical variable 2

For CV 2, two figures are present. Each figure shows the two-dimension embedded location of the 500+ features of CV 2. As with CV 1, the colour of the feature represents the cluster it belongs to, this is where the two figures differ. Figure 10 uses the clusters made by the experts at the insurance company. Figure 11 shows the 27 clusters made using the K-means method based on the location in the non-reduced embedded space for CV 2.

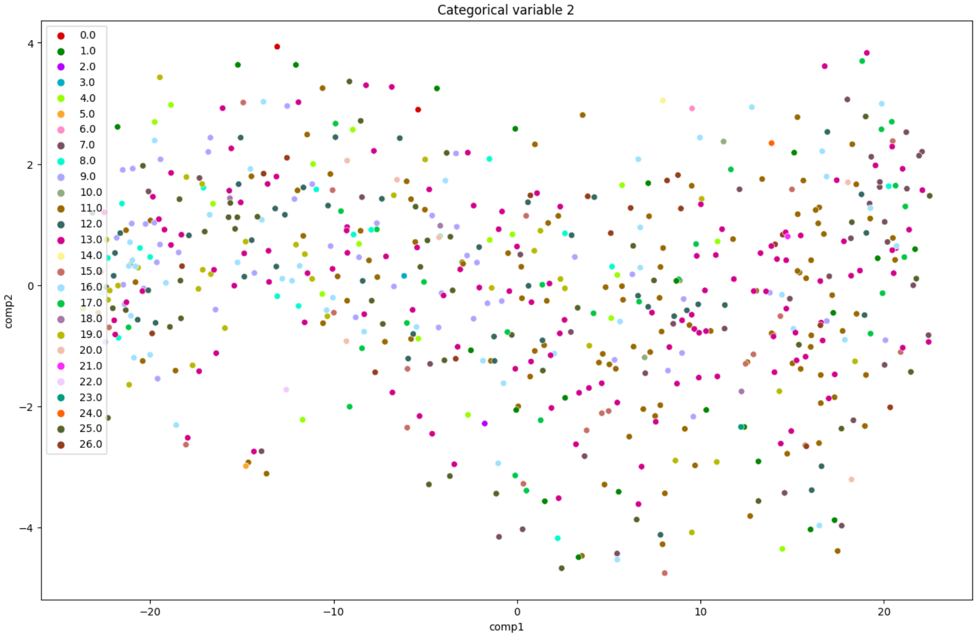


Figure 10: Location of variables in CV 2 in a reduced embedded space. Clustered using the insurance companies clustering.

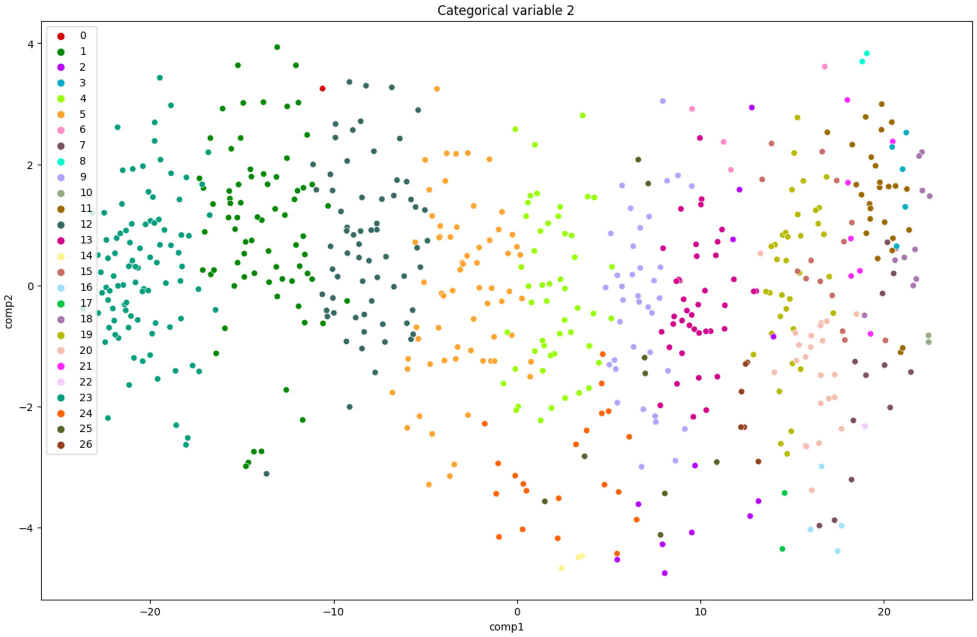


Figure 11: Location of variables in CV 2 in a reduced embedded space. Clustered using the K-means method.

## 4.4 Impact of clusters

In the previous sections, some results of the clustering were shown. As mentioned in Chapter 3, all these clusters are individually tested alongside the original risk factors of the insurance company in the frequency GLM.

In Figures 12, 13, 14, 15 the mean Poisson deviance of the different GLMs on the test set is plotted. The y-axis shows the mean Poisson deviance. The x-axis represents the number of clusters made using the chosen cluster method. As each numerical variable was clustered using four different methods, four lines are plotted in each figure. Furthermore, the mean Poisson deviance of the base GLM model is also plotted in each figure, this is represented by the black line.

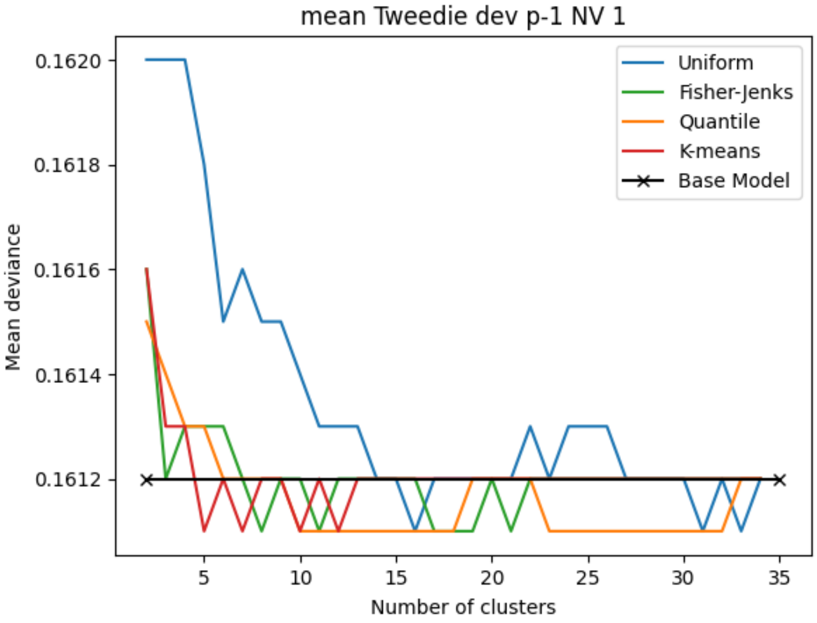


Figure 12: Mean Poisson deviance of the GLM with NV 1 clusters.

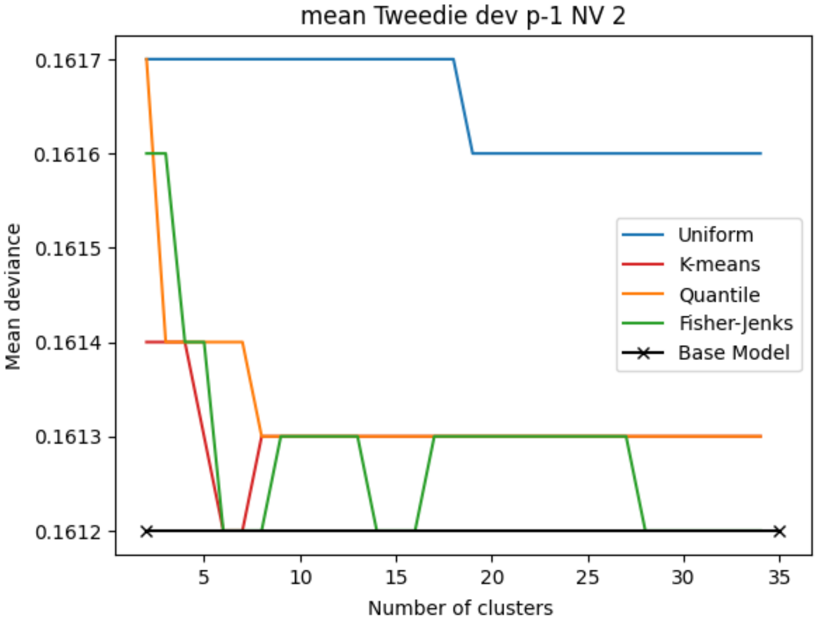


Figure 13: Mean Poisson deviance of the GLM with NV 2 clusters.

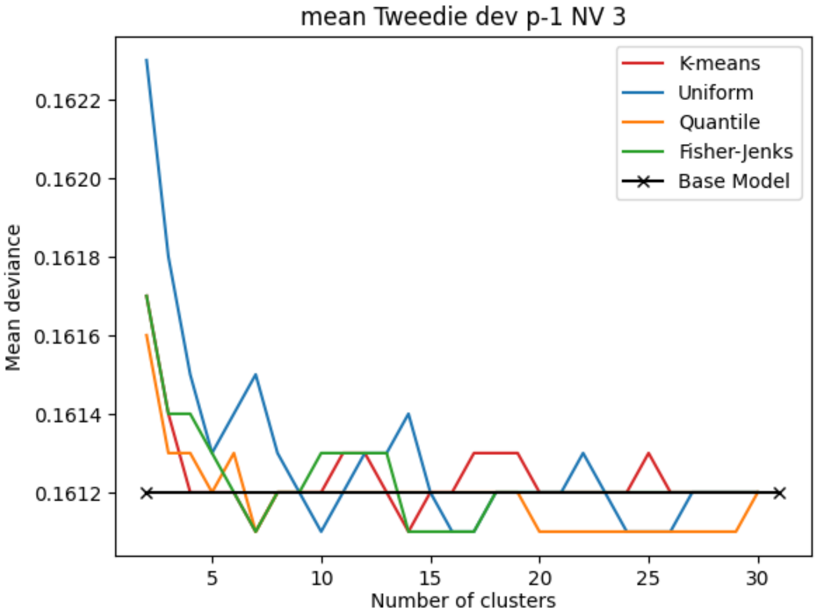


Figure 14: Mean Poisson deviance of the GLM with NV 3 clusters.

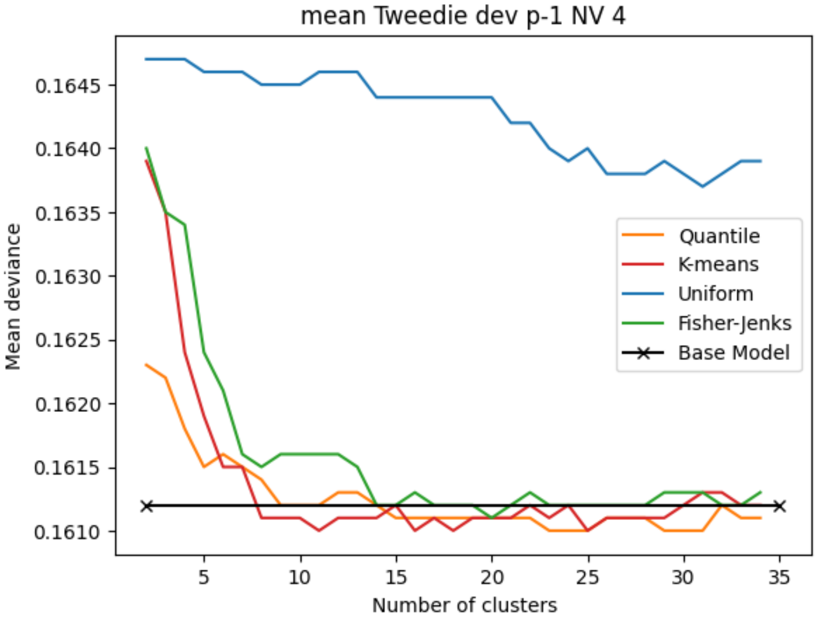


Figure 15: Mean Poisson deviance of the GLM with NV 4 clusters.

As with the clustering results, only the results of the NN with a hidden layer size of 10, trained using continuous inputs, are shown in the figures below. The results of the NN with the hidden layer size of 120 trained on categorical inputs and the results of the NN with the hidden layer size of 120 trained on the continuous variables are shown in Appendix 7.1.

As with the previous figures for the numerical variables, Figure 16 and 17 show the Mean Poisson deviance of the GLM in which the only one variable was changed compared to the base mode. Figure 16 contains the results of the GLM in which CV 1 was replaced and Figure 17 the results in which CV 2 was replaced. Both figures also contain the black line that represents the Mean Poisson deviance of the base GLM.

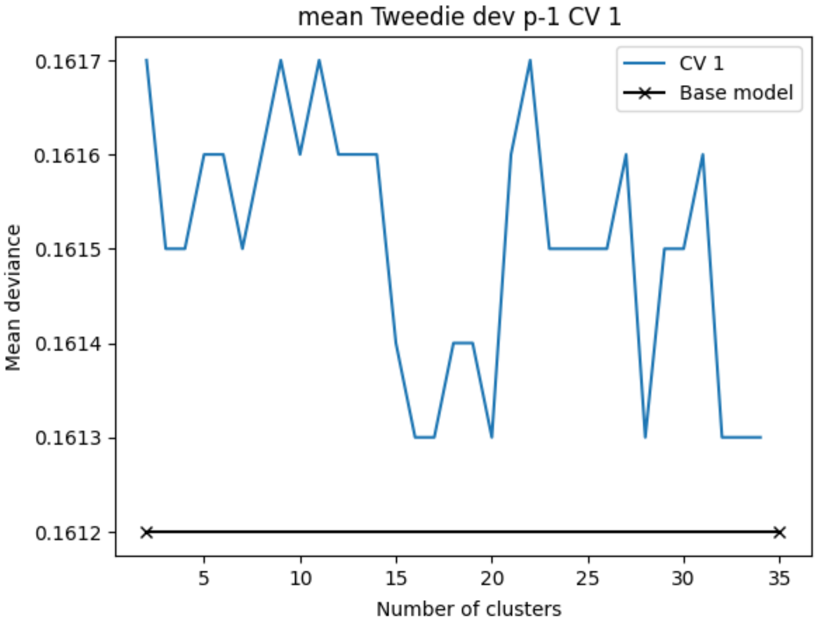


Figure 16: Mean Poisson deviance of the GLM with CV 1 clusters.

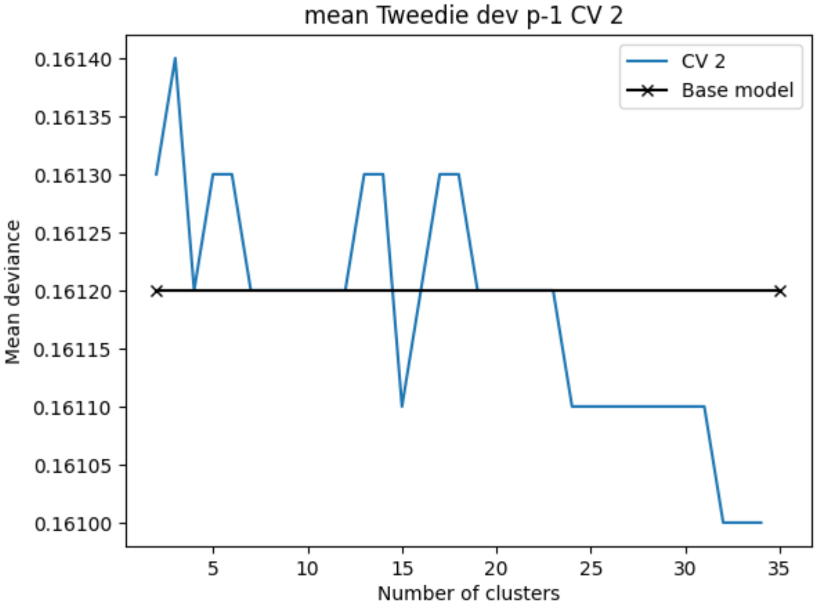


Figure 17: Mean Poisson deviance of the GLM with CV 1 clusters.

## 4.5 Impact of clusters on the pure premium

In Chapter 3, it was also mentioned that GLM models would be created using more than one new cluster. In the tables below, the results of the different GLMs are depicted. Each table contains the results of the base mode. All NV and CV variables changed, changing all numerical variables, changing of only CV 1, changing of only CV 2 and a hand-picked GLM. Table 4 contains the NN with categorical input. Table 5 contains the NN with continues inputs with hidden layer size of 10 and Table 6 the NN with hidden layer size of 120. The different NN clustering methods have no influence on the base GLM or the numeric GLM.

The number of clusters chosen for the numerical risk factors depends on the elbow method. For all numerical variables, the Jenks method provided the best result. NV 1 was decided to have 8 clusters, NV 2: 7 clusters, NV 3: 7 clusters and NV 4: 9 clusters. Both CV 1 and 2 were selected to have 27 clusters. For the hand-picked method, the numerical variables were chosen based on the deviances found in Section 4.4. Here NV 1, 2 and 3 still used clusters found by the Jenks method, the chosen sizes are 18, 7 and 15 respectively. For NV 4 it was chosen to use the K-means method with a size of 17 clusters. To select the size of the categorical variables, the deviance plots were also used. CV 2 stayed the same amount of cluster at 27, we did not want to go over the cluster size used by the insurance company. For CV 1 it is chosen to use the original non-clustered variable.

Each table below shows the D$^2$, the mean Poisson deviance, mean average error and mean squared error. Further down, the pure premium results are shown. These were obtained by multiplying the results of the various GLMs on the test set with the results of a base severity GLM on the same test set. The shown pure premium results are the mean, standard deviation, the minimum premium value and the maximum premium value.

| | Base GLM | All new | Numeric | CV 1 | CV 2 | Picked |
|---|---|---|---|---|---|---|
| Dev$^2$ exp | 0.0744 | 0.0735 | 0.0733 | 0.0735 | 0.0757 | 0.0770 |
| Dev | 0.1612 | 0.1614 | 0.1614 | 0.1614 | 0.1610 | 0.1608 |
| MAE | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 |
| MSE | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 |
| Mean PP (€) | 76.83 | 76.60 | 76.68 | 76.80 | 76.76 | 76.50 |
| Std PP (€) | 62.81 | 66.39 | 66.26 | 62.15 | 63.45 | 65.12 |
| Min PP (€) | 2.56 | 2.16 | 2.36 | 2.76 | 2.35 | 2.44 |
| Max PP (€) | 1049.43 | 1175.34 | 1276.20 | 963.98 | 1054.95 | 1348.6 |

Table 4: GLM results and effects on the pure premium. NN Categorical input data.

| | Base GLM | All new | Numeric | CV 1 | CV 2 | Picked |
|---|---|---|---|---|---|---|
| Dev$^2$ exp | 0.0744 | 0.0722 | 0.0733 | 0.0724 | 0.0751 | 0.0766 |
| Dev | 0.1612 | 0.1616 | 0.1614 | 0.1616 | 0.1611 | 0.1609 |
| MAE | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 |
| MSE | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 |
| Mean PP (€) | 76.83 | 76.35 | 76.68 | 76.65 | 76.71 | 76.45 |
| Std PP (€) | 62.81 | 66.23 | 66.26 | 61.33 | 64.98 | 66.49 |
| Min PP (€) | 2.56 | 2.06 | 2.36 | 2.74 | 2.47 | 1.85 |
| Max PP (€) | 1049.43 | 825.34 | 1276.20 | 914.15 | 997.82 | 1223.79 |

Table 5: GLM results and effects on the pure premium. NN 10.

| | Base GLM | All new | Numeric | CV 1 | CV 2 | Picked |
|---|---|---|---|---|---|---|
| Dev$^2$ exp | 0.0744 | 0.0738 | 0.0733 | 0.0741 | 0.0752 | 0.0766 |
| Dev | 0.1612 | 0.1614 | 0.1614 | 0.1613 | 0.1611 | 0.1609 |
| MAE | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 | 0.0386 |
| MSE | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 | 0.0408 |
| Mean PP (€) | 76.83 | 76.26 | 76.68 | 76.82 | 76.75 | 76.49 |
| Std PP (€) | 62.81 | 67.95 | 66.26 | 63.28 | 65.04 | 66.52 |
| Min PP (€) | 2.56 | 2.07 | 2.36 | 2.43 | 2.36 | 1.91 |
| Max PP (€) | 1049.43 | 1085.31 | 1276.20 | 1100.94 | 994.64 | 1193.13 |

Table 6: GLM results and effects on the pure premium. NN 120.

### 4.6 Comparing the base model with the picked model

In the following section, the base model is compared with the picked model of the NN with categorical input data. First, the cumulative distribution of the weights for a risk factor is shown. The y-axis of the plot shows the weight of a feature in the risk factor. The x-axis shows the cumulative presence of the features in the risk factor. The jump a new feature makes in the plot shows how much it is present in the test data for the GLM. All figures contain two plots, one for the base mode and one for the picked model, their respective plots are indicated in the legend. Only risk factor NV 1, NV 4, CV 1 and CV 2 are shown in Figure 18, 19, 20 and 21 respectively. All the other figures can be found in Appendix 7.3.

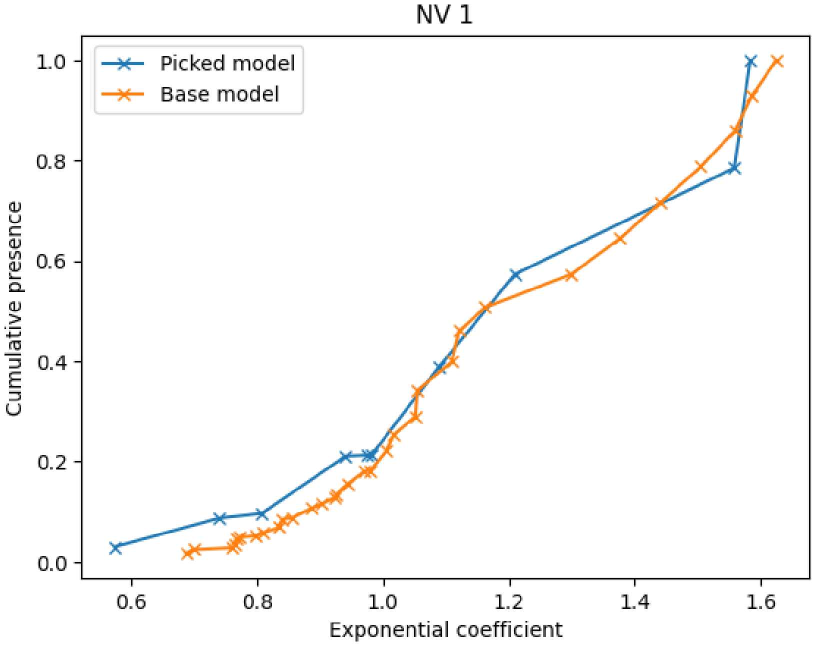


Figure 18: Cumulative distribution of the weights of NV 1 for the GLMs.

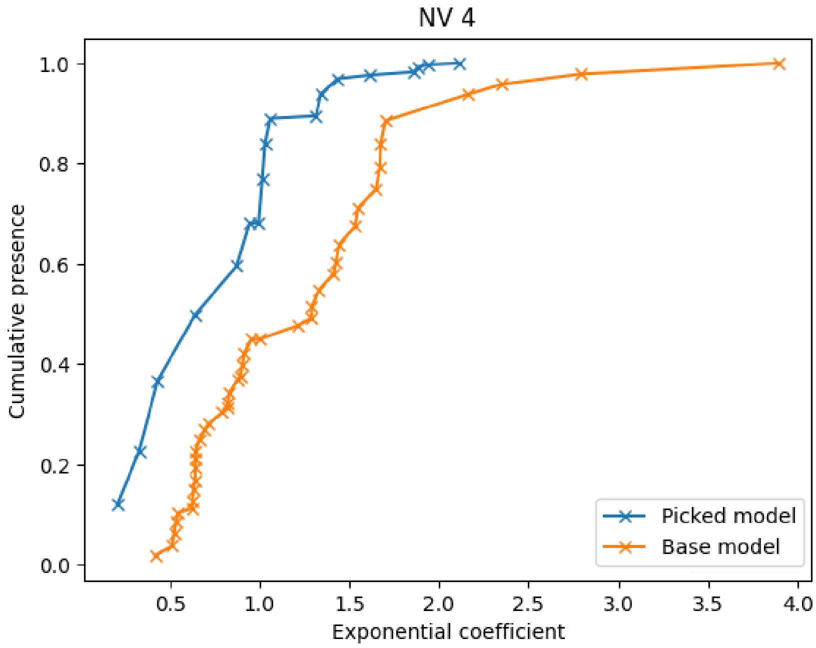


Figure 19: Cumulative distribution of the weights of NV 4 for the GLMs.

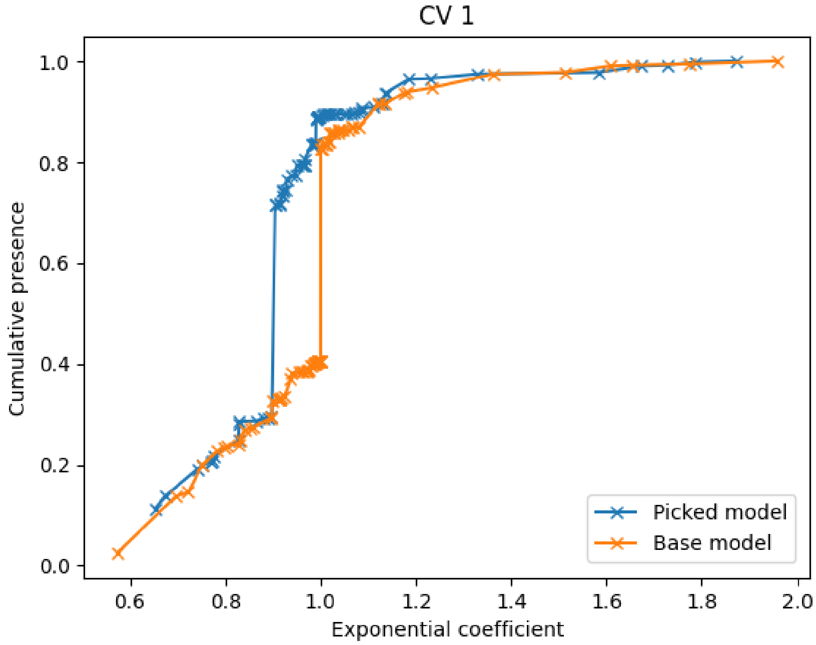


Figure 20: Cumulative distribution of the weights of CV 1 for the GLMs.

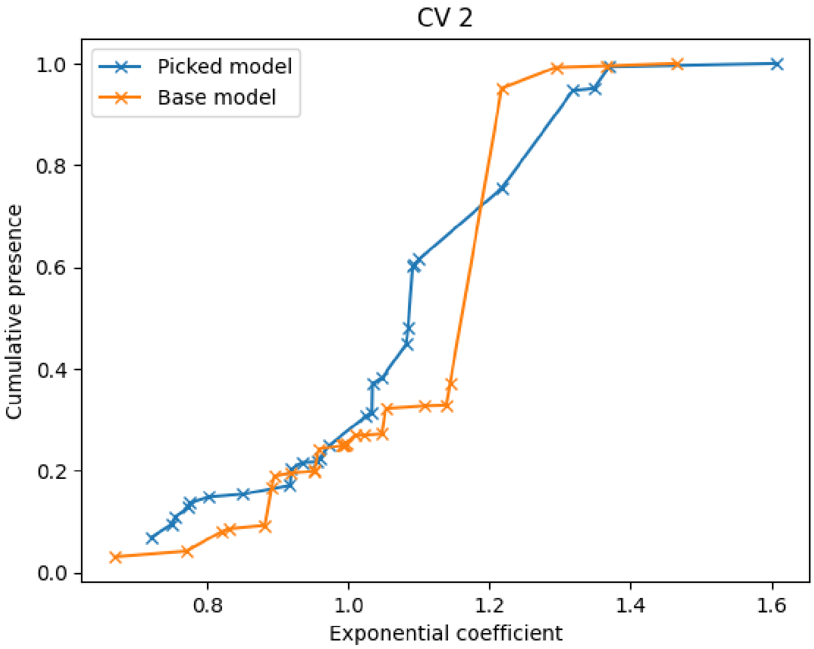


Figure 21: Cumulative distribution of the weights of CV 2 for the GLMs.

### 4.6.1 Difference in pure premium

The difference between individual premiums when computed with the base model and picked model is found by subtracting the premium of the picked model from the premium of the base model. Figure 22 shows a histogram plot of the differences in euros (€). The x-axis shows how much the premiums differ in euros (€). The y-axis shows how many premiums differ for a given difference.

The previously found difference in premium price is used to find the percentage change of the pure premium. The percentage change is found using Equation 23. In Figure 23 the percentage change of the pure premium is plotted. The x-axis shows how much the picked model premiums change compared to the base model premiums in percentage. The y-axis shows how many premiums differ for a given percentage difference.

$$PP_{\%} = \frac{PP_{pm} - PP_{base}}{PP_{base}} * 100 \tag{23}$$

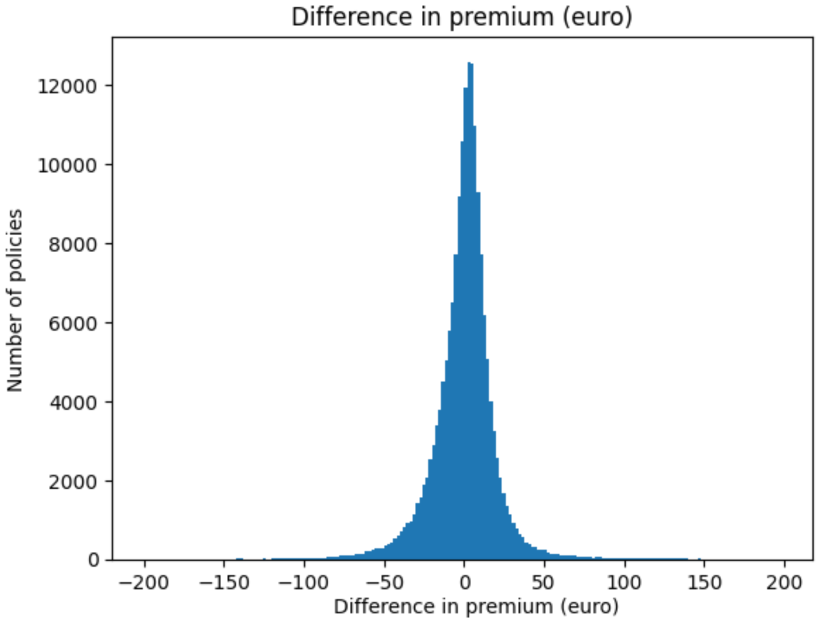


Figure 22: Difference in premium price between the base model and the picked model.

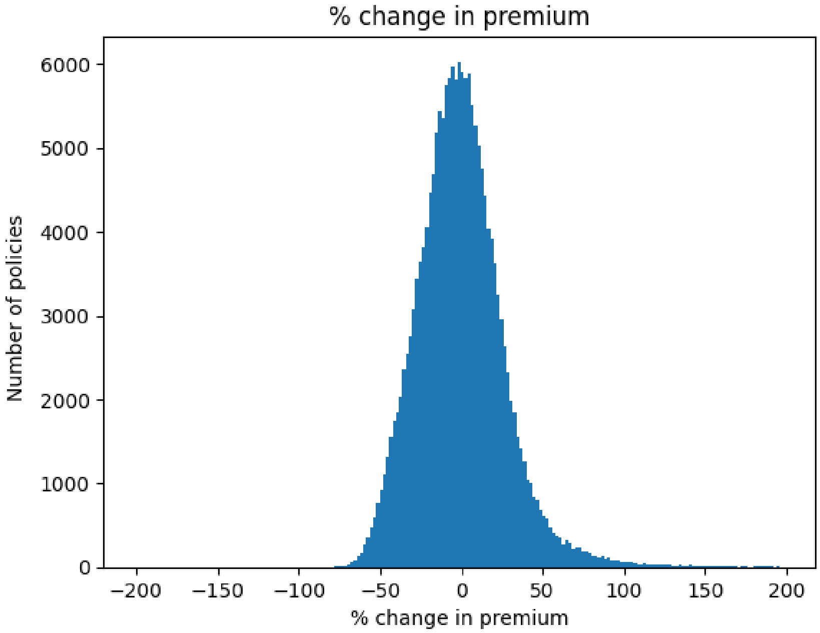


Figure 23: Percentage change in the pure premium of the picked model when compared to the base model.

31.53% of the policies in the test set saw a change in pure premium bigger than 25%. 5.73% of the policies in the test set saw a change in pure premium bigger than 50%. This change in premium percentage takes both an increase and decrease into account.

The policies whose premiums change with more than 50% can be split into two groups, increased and decreased premiums. 2.19% of the policies in the test set saw a pure premium reduction of more than 50% while 3.53% saw an increase of more than 50%.

The average weights of the risk factors for both of the 50% groups is found for both the base GLM and the picked model GLM. The mean weights of the GLM models and their differences for the group in which the base model pure premium is cheaper is found in Table 7. Table 8 also shows this, but for the policies in which the base pure premium was more expensive.

| | Base | Picked | Diff |
|---|---|---|---|
| IV 1 | 1.0216 | 1.0200 | -0.0016 |
| IV 2 | 0.9967 | 0.9930 | -0.0037 |
| IV 3 | 1.0272 | 1.0299 | 0.0027 |
| CV 1 | 0.8505 | 0.8836 | -0.0331 |
| CV 2 | 1.2148 | 0.8663 | 0.3485 |
| NV 1 | 1.3206 | 1.2524 | 0.0682 |
| NV 2 | 1.3515 | 1.2189 | 0.1326 |
| NV 3 | 1.1811 | 1.1722 | 0.0089 |
| NV 4 | 0.7470 | 1.1474 | -0.4004 |

Table 7: Mean weights of the GLM models and their differences for premiums in which the base pure premium is at least 50% cheaper than the picked model

| | Base | Picked | Diff |
|---|---|---|---|
| IV 1 | 1.0256 | 1.0237 | 0.0019 |
| IV 2 | 1.0156 | 1.0042 | 0.0114 |
| IV 3 | 0.9841 | 1.0021 | -0.0180 |
| CV 1 | 0.9239 | 0.9146 | 0.0093 |
| CV 2 | 1.1681 | 0.8179 | 0.3502 |
| NV 1 | 1.0729 | 0.9928 | 0.0801 |
| NV 2 | 1.4142 | 1.5521 | -0.1379 |
| NV 3 | 1.0210 | 0.9621 | 0.0589 |
| NV 4 | 0.7540 | 0.3187 | 0.4353 |

Table 8: Mean weights of the GLM models and their differences for premiums in which the base pure premium is at least 50% more expensive than the picked model

## 4.7 ML models to replace the frequency GLM

As mentioned in Chapter 3, three different types of ML models are tested in their performance of modelling the claim frequency. These ML models are tested based on the existing cluster used by the insurance company. For each of the models, the mean squared error and mean average error over the test set are found.

The found claim frequencies on the test set are multiplied with the results of the base severity model to get an indication of the pure premium when using ML models. Table 9 contains the error results and pure premium results of the models using the existing cluster.

| | XGB Cat | NN Cat | RF Cat |
|---|---|---|---|
| MSE | 0.1395878 | 0.1397136 | 0.1401506 |
| MAE | 0.0426750 | 0.0457891 | 0.0418469 |
| PP Mean | 78.96 | 90.33 | 75.58 |
| PP STD | 72.21 | 25.35 | 106.25 |
| PP MIN | -65.63 | -5.49 | 31.94 |
| PP MAX | 1398.64 | 136.71 | 21862.67 |

Table 9: Results of the ML models using the one-hot encoded data of the GLMs.

# 5 Discussion

In this chapter, we discuss the results in more detail. We find the conclusion to the research questions and interpret the results. Furthermore, we discuss the drawbacks of this project and outline considerations for future research.

## 5.1 Answering the research question

In the introduction, we presented the research goal, which is as follows:

”To understand the extent to which clustering within variables improves the current modelling method, of using a GLM, for determining the pure premium. Furthermore, we look at how the GLM modelling technique compares to other more novel and future forward machine learning modelling techniques for the pure premium when using clustered data.”

For this thesis, we performed multiple experiments to help answer the research goal and the corresponding research questions. These sub research questions are shown below and are also found in Chapter 1.4.

1. Which are the current methods to handle numerical and categorical data for GLMs?
2. How can clusters be made, along with determining their optimal number, to ensure that each cluster significantly influences the pure premium?
3. How do combinations and variations of numerical and categorical clustering impact the pure premium?
4. How does pure premium modelled by the GLM compare to novel modelling techniques?

We answered the first research question by looking into the different ways GLMs can handle categorical and numerical data. This information is found in the background chapter of this thesis (Chapter 2) and answers research question one. We found that for categorical data, entity embedding can be used to cluster in a data-driven way such that the GLM can handle the data. For numerical data we found the following clustering methods: Uniform, Quantile, Fisher-Jenks and K-means.

Research question two was answered with the experiments introduced in Chapter 3. These experiments looked at multiple different clustering techniques, some applicable to numerical variables, others to categorical variables. By examining the goodness of fit and deviance score of the GLM on the test set, we identified the optimal cluster amounts per variable. The plots of the deviance showed that changing the clustering size impacted the deviance of the GLM model up to a certain amount of clusters. After a certain amount of clusters, the deviance did not improve any more.

To answer the third research question, we tested multiple combinations of newly clustered risk factors and existing clusters, as explained in Section 3.5. These combinations showed that each model made looked almost identical when compared to each other based on error rates, mean pure premium and the standard deviation of this pure premium. However, a closer look into the individual premiums shows that larger differences are present. Thus, the answer to research question three is that the impact of clustering on the whole policy level of the premium is small, but the impact on individual premiums can be large.

We answered the final research question by creating multiple ML models, these models are described in more detail in Chapter 3. These models showed that when using the data in a one-hot encoded state, the GLM performed the best when looking at the errors. The models also showed that some more novel modelling techniques provide impossible real world answers in this experiment, such as negative pure premiums. The GLM and random forest, one of the novel modelling techniques, provided plausible real world values for the pure premium.

## 5.2 Interpretation of the results

In the following section, we give a more detailed observation of the results. First, we discuss the results for the numeric clustering, and then we follow up with the results of the categorical clustering. Then we look at the impact of these clusters on the GLM in more detail. Afterwards, the results of the different combinations of risk factors are discussed. Then we compare two models in more detail on why and how the premium would differ between them. Finally, we compare the results of the ML models to the GLMs.

### 5.2.1 Numerical clustering

The first results for the numerical risk factor clustering were found by plotting how well the different clustering techniques and sizes capture the variability of the risk factors data set. Each of the plots in Figure 5, 6, 7 and 8 represent the results of this experiment for the variables NV1, NV2, NV3 and NV4 respectively. In all four of the figures, the Fisher-Jenks algorithm represented by the green line always has the highest goodness of fit for a given cluster amount. This indicates that the Fisher-Jenks clusters explain the variance of the data better than the other methods. Furthermore, it can be seen that increasing the number of clusters also increases how much variance is explained when using this clustering method. This is seen by the increasing green line with an increase in cluster amount in the figures.

Looking specifically at NV 1, in Figure 5, it can be seen that the K-means clustering method closely follows the trajectory of the Fisher-Jenks clustering method. The other two clustering methods, Uniform and Quantile, clearly explain less of the variance for a lower amount of clusters. This is made clear by their lines in the goodness of fit plot, which are lower than the other lines of K-means and Fisher-Jenks clustering method. This difference, in how well the different methods explain the variance, is smaller as the amount of clusters increase. An increase in the number of clusters for these methods, similar to the Fisher-Jenks method, demonstrates that more of the variance is explained.

For NV 2, 3 and 4, in Figures 6, 7 and 8, the same trend can be seen for the Uniform and Quantile methods. This means for a lower amount of clusters, these methods work less well at explaining the variance when compared to the Fisher-Jenks method. As mentioned previously for the NV 1, the K-means method creates clusters that explain the variance well, this also holds for NV 3. However, this method clearly falls behind the Fisher-Jenks method for NV 2 and 4. This further indicates that the Fisher-Jenks clustering method is the optimal method of the four tested numerical variables for this data set when testing on the explained variance.

The optimal number of clusters was chosen based on the elbow method as described in Chapter 3. For NV 1 it was chosen to have 8 clusters, NV 2: 7 clusters, NV 3: 7 clusters and NV 4: 9 clusters. All of these are chosen based on the Fisher-Jenks clustering method. A downside of the method is that the elbow point is subjective.

### 5.2.2 Categorical clustering

The results of the entity embedding were visualized in Chapter 4.3. As mentioned in Chapter 2, features with similar effects on the model are located close to each other in the embedded space. The dimensional reduction for the visualization tries to keep the similar features close together.

Figure 9 shows the position of each feature found in risk factor CV 1 in the reduced two-dimensional space. What can be seen from this figure is that the dimensional reduction step does not always preserve the closeness of the features. This can be seen by looking at the lime green dots in the figure. These dots represent the Cluster 4 (made in the embedded space). Three of these features are located together at around -4 and -1.5 for the comp 1 and comp 2 axes respectively, however one of the points is located at -5, -2. This feature would be placed in a different cluster if the cluster was formed after the dimensional reduction. This shows that the reduction step does not always preserve the closeness of the features. Figure 9 also shows that the number of features in a cluster can vary greatly between clusters. This is evident from comparing the amount of features in Clusters 2 and 22.

The results of the clustering for CV2 can be found in Figure 10 and 11. Here, Figure 10 shows the clustering of the insurance company and Figure 11 shows the clustering based on the entity embedding. Both of these figures show the location of features, just with different clustering. Figure 10 clearly shows that the features in a cluster are not located close to each other for almost all clusters. Looking at the features in Clusters 8 and 13, it can be seen that features are spread widely apart from one another. This indicates that the clusters made by the insurance company did not contain features that had a similar effect on the trained NN model that predicts the frequency of the pure premium. The clusters made by the insurance company are based on explainability. Features that logically belong together are placed in the same cluster.

Figure 11 shows the same image, but now with the clusters formed based on the location in the embedded space. It can be clearly seen that the features in each cluster are located more closely together compared to the insurance company's cluster. As with CV 1, the number of features in a cluster varies greatly, this can be seen by comparing Cluster 0 with Cluster 23 in Figure 11. Likewise, the upper right side of the figure shows that the dimensional reduction again does not always preserve the closeness of the features from the embedded space.

One of the main issues with the clustering via entity embedding is that clusters can contain features that might not seem to fit together, this reduces the explainability. We can confirm that the embedded clusters have a less explainable combination of features by comparing the two figures. Looking on the right side of Figure 11 we can see numerous features belonging to one cluster, Cluster 23. These same features belong to many different clusters when looking at Figure 10, the insurance companies clustering method. This shows that the embedded cluster contains features that logically might not go together according to the insurance companies clustering method.

### 5.2.3 Impact of the clusters

How the numerical clusters impact the base GLM is shown in Figures 12, 13, 14 and 15. A lower mean deviance indicates that the model performs better on the test set. As mentioned previously, this is also the metric used by the insurance company to evaluate the GLMs.

In all the figures it can be seen that the blue line, indicating the Uniform clusters, starts with the highest deviance. For NV 2 and 4 this line does not reach the same level of deviance as the other clustering methods. For NV 1 and NV 3, it reaches a stable level later than other methods, however for NV 3 this difference is smaller than for NV 1. This indicates that the distribution of the numerical risk factors plays a part in how well a clustering technique performs. Furthermore, based on these plots it can be concluded that the Uniform clustering method performs the worst, in general.

Looking at the other clustering methods, it can be seen that a higher explained variance of clusters does not always mean a lower deviance in the GLM. This can be seen when looking at the Fisher-Jenks plot in Figure 8 and Figure 15. In Figure 8, it can be seen that this method has the highest explained variance for all given cluster amounts. However, in Figure 15 the Fisher-Jenks clustering method has a higher mean deviance compared to the K-means and Quantile method.

The GLM simulates the relationship between the data and the targets via the input clusters. A different cluster structure, as in the placement of the cluster bins, affects how the GLM interprets and models the relationships. The higher variance explaining clusters, made with the Fisher-Jenks method, might capture more detailed patterns in the data but not necessarily those relevant to the claim frequency.

From the four figures, it can also be seen that the deviance of all GLMs starts lowering when the amount of clusters increase. However, for all numeric risk factors and all clustering methods, except for the Uniform method, the decrease in deviance levels out before 15 clusters. When the lines in the plots have stopped decreasing, the mean deviance deviates with roughly 0.001 around a set deviance value. The stagnation of the deviance can also be seen in the research of Henckaerts et al., 2018.

What the figures also show is that none of the cluster methods leads to a GLM that reaches a mean deviance far lower than the insurance company's GLM. The lowest mean deviance was 0.002 lower than the base model. This can be partly explained by the fact that between these models only one risk factor is clustered differently.

From the deviance plots it can be concluded that for this data set all clustering methods, except for Uniform clustering, produce very comparable results to each other. These results are also comparable to the base model results, given a sufficient number of clusters is used for the data-driven clustering methods. This indicates that the optimal cluster method and number of clusters differ based on the assessment criteria of explained variance and mean deviance.

As mentioned in Chapter 3, the insurance company uses the Poisson deviance as one of the criteria to determine what GLM model they use. While some clusters do make a GLM with a lower deviance than the base model, the difference is small. This shows that data-driven clustering only marginally improves the frequency GLM when changing only one of the risk factors clusters.

The impact of replacing one of the categorical risk factors CV 1 and CV 2 is shown in Figures 16 and 17 respectively. Figure 16 shows that clustering risk factor CV 1 increases the mean deviance of the frequency GLM. This indicates that clustering categorical variables does not always lead to a better model. Figure 16 also shows that clustering the high cardinal categorical risk factor is also sensitive to the amount of clusters used. An increase in clusters leads to a decrease in mean deviance of the GLM.

Looking at a cluster amount of 27 on the x-axis in Figure 17 we observe that the embedded clustering has a very slightly lower deviance than the clustering used by the insurance company (also using 27 clusters). However, based on the difference in the cluster location of Figures 10 and 11 it was expected that the new clusters would have a bigger impact on the deviance of the GLM. The small decrease in mean Poisson deviance compared to the expert's clustering does not seem to be worth the loss of explainability.

### 5.2.4 Impact of clustering on the pure premium

To answer the third research question, different combinations of the previously found optimal clusters for the various risk factors were combined to create multiple GLM models. Due to the different methods of training, the NN three tables were made. This also allows for the comparison between the different NN training methods.

Looking at Table 4, the first thing to note is the MAE and MSE between all models is the same. This indicates that the differences in the performance of the different models are not large. Looking at the deviance, a difference in the models can be seen. We expect to see a larger change in deviance between the models than a change in MSE and MAE. This occurs because the deviance returns a larger error for the same measurements due to the nature of how the deviance is calculated. The MSE and MAE having no difference after the fourth decimal place is also observed by de Bont, 2022. Here, the models with the same risk factors only showed a difference at the sixth decimal place.

The model that changed all the numeric risk factors with the Fisher-Jenks optimal had a worse mean deviance than the base model. Thus, using this method of finding optimal clusters does not produce a better model in this instance. This result was to be expected based on the results of Section 4.4. As, changing the individual clusters does not decrease the deviance below the base model. Especially, the deviance of NV4 was higher than that of the base model for the chosen cluster amount.

The model that changed all numeric and categorical risk factors also performed worse than the base model. This is also expected when seeing that replacing all the clusters for the numerical risk factors resulted in a higher deviance. This was also observed when only changing the clusters of CV 1.

The two models that did decrease the mean deviance is the model with hand-picked replacements and the GLM that only replaced CV 2 and is a repeat of the results seen in Figure 17. The hand-picked replacements are based on the best deviance's found by the individual replacements in the GLM (the results of Section 4.4). From this, it can be concluded that picking the optimal clustering method and amount for this data set is based on the cluster that achieved the lowest mean Poisson deviance when the risk factors were replaced individually.

The lower half of Table 4 shows the effects of the model on the pure premium. The difference in the average pure premium between all the models in the table ranges from €0.03 to €0.33. The difference in standard deviation between the models is slightly larger, ranging from €0.13 to €4.28, this indicates that the spread of premiums does change more than the average premium between models. This is further supported by the difference seen in the maximum pure premium for the given models. The difference in the minimum pure premiums is relatively small.

Looking at Table 5 and 6 reveals the same conclusions as those found for Table 4. However, there are differences between the tables. The clustering model that used a NN with a hidden layer size of 10 produced the highest mean deviance GLMs for both the CV 1 and CV 2 replacement. None of the models replacing CV 1 had a lower mean deviance than the base model.

The tables also show that the picked model in Tables 4, 5 and 6 all have a lower mean deviance than the base model. The lowest mean deviance can be found in the picked model of Table 4. This indicates that choosing a combination of clusters that individually produced the lowest deviance produced a better model than the choosing cluster based on the elbow method on the variance explained plots.

### 5.2.5 Comparing models

The base model and the picked model from Table 4 were compared to give more insight into the differences between the models. Figure 18 shows the base model uses more clusters for the risk factor NV 1. Each cluster is represented by $x$ and the base model plot has more clusters than the picked model plot. The higher amount of clusters leads to more possible coefficients, giving the possibility for the model to better fit the data, but this also contributes to a more complex model.

Figure 18 also shows that the coefficients of the picked models start and end lower than the base models coefficients. Overall, the distribution of the coefficients for the NV 1 clusters follows a similar path between the two models. One difference is that the picked model clearly starts with a lower coefficient, this might contribute to the lower minimum premium this model has.

Figure 19 shows a clear difference in the distribution of the coefficients over the data. Over 60% of the policies in the test set have a coefficient less than 1 for the picked model, whereas this is around only 40% for the base model. This shows that the picked model, on average, has a lower coefficient for NV 4 than the base model. Furthermore, the maximum coefficient for this risk factor is also larger for the base model than for the picked model.

The distribution of the coefficients over the data for CV 1, seen in Figure 20, differs from the other risk factors as the same clusters are used in both models. However, it can be seen from the figure that the plots are not the same. This shows that the GLM has given new coefficients to the CV 1 features when trained with the new cluster for the other risk factors. This can clearly be seen by the different location of the vertical line in the two plots. The vertical line indicates that a large portion of the data belongs to the cluster with coefficient at the endpoint of that line (the upper $x$ of the vertical line). The figures show that this cluster, identical for both GLMs, has different coefficients for the two GLMs.

From Figure 21, we can see that the base model has one cluster that is present in around half of the policies of the test set. This can be seen from the large increase in cumulative presence from 0.4 to more than 0.9. For the new clusters used by the picked model, the policies are divided more evenly, with no cluster having more than 20% of the data points. The spread of the coefficient for both models is comparable to that of NV 1.

These images show that clustering can alter the representation of relationships between variables, as seen by the changes in the coefficients. The difference in the amount of clusters, and thus edges of the clusters, can lead to coarser or finer clusters between the different models. These clusters may emphasize different trends or overlook important details found in the original continuous data.

### 5.2.6 Comparing the pure premium between the models

In the previous section, we discussed the differences between the distribution of the coefficients for two chosen models. In this section, we discuss the results of how the pure premium differs between the models, thereby fully answering the third research question.

Figure 22 shows the difference in pure premium between the models for all policies in the test set. This plot shows that the change in premiums is concentrated around €0, with roughly an equal amount of premiums increasing as decreasing. This is to be expected, as the average premium of each model only differed €0.33. However, at the extremes, the pure premiums can change with more than €100 for a policy.

The percentage change of the pure premium if the policies would switch from the base model to the picked model is shown in Figure 23. As with the exact change in euros, the percentage change is centred around 0%. As mentioned in Chapter 4, 31.53% of the policies in the test set saw a change in pure premium bigger than 25%. This is a large change in pure premium for a model that only slightly improved the mean Poisson deviance, while not improving the MSE or MAE.

Henckaerts et al., 2018 and Henckaerts et al., 2021 both show that when the difference in average premium between two different models is very small, the individual premiums can still differ a lot between the models. Furthermore, the shape of the relative difference in premiums matches that of Figure 23. However, the percentage change in Henckaerts et al., 2018 and Henckaerts et al., 2021 are more strongly centred around 0 and thus have a lower number of policies experiencing big changes between the models.

Changes in individual premiums are to be expected when the clustering strategy between the models differ. We saw earlier how clustering can alter the representation of relationships between variables in the GLM models. Furthermore, variables of an individual premium might be clustered differently between the two techniques and as different clusters have different coefficients, a different premium is to be expected. Figure 23 shows that this difference can increase or decrease a premium. The difference between the two selected models do balance out, as seen by the similar average premiums and the centred plots around 0 in Figure 22 and 23.

### 5.2.7 What caused the largest changes in pure premium

Tables 7 and 8 show the average risk factors coefficients of the policies that are a lot cheaper and a lot more expensive. What can be seen from both these tables is that the biggest differences between the two models can be found in the risk factors of CV 2, N2 and N4.

The coefficients for CV 2 stay relatively the same between the two tables, as does their difference. This could indicate that the large change in premiums is not driven by this risk factor, as it is present in both an increase and a decrease in premium.

The coefficient for the risk factor NV 2 is 0.133 higher for the base model compared to the picked model when the base model is cheaper, as seen in Table 7. Based on Equation 11, it is expected that the cheaper model would have a lower coefficient. When the base model is more expensive, the difference goes to $-0.138$, again opposite of what is expected. This would indicate that NV 2 is not the risk factor that drives the large differences in the pure premium.

The expected change in coefficients can be found in NV 4. Here, the coefficient for the base model is 0.40 lower when this model is cheaper and 0.43 higher when this model is more expensive. This change can be fully attributed to the picked model, as the average coefficient of the base model does not change much between Table 7 and 8. From this, it can be concluded that NV 4 plays a big part in the biggest percentage changes of the pure premium between the models.

### 5.2.8 ML models to replace the frequency GLM

As mentioned in Chapter 2, different methods exist for modelling the frequency component of the pure premium. Table 9 shows the results of these models with the same input the GLM received.

Looking at the MSE and MAE of the different models in Table 9, it can be seen that these are higher than the GLM models found in Section 4.5. This is the first indication that using one-hot encoded data does not work well for these models when predicting the frequency component of the pure premium for this research. The second indication can be seen in the minimum values of the pure premium. Here, the XGB and the NN models produce negative values. This is due to negative prediction on the frequency of a claim, something which is not possible in the real world. Based on error values and the negative pure premium, the XGB and NN models can not replace the GLM when using one-hot encoded data. The RF does not produce impossible pure premium numbers, however the maximum pure premium is unrealistically large and would probably not be accepted by the insurance company and the policyholder. This combined with the larger error also makes the RF model unsuitable to replace the GLM when using the same input.

## 5.3 Drawbacks and limitations

As with any research, this project has its drawbacks and limitations. In this section, we highlight and discus these in more detail.

### 5.3.1 Data preparation

While a few instances of wrong policies have been removed, there still exists strange policies, missing data and outliers for certain values in the data (Super high insured value). This could impact the clustering techniques and thus the model fitting. Furthermore, the chosen split of training and testing data could have also influenced the results of the training.

### 5.3.2 Numeric clustering

For the numeric clustering, the elbow point is subjective and can differ based on the person analysing the same data. Ketchen and Shook, 1996 states that in many applications, the elbow point is highly ambiguous. This is also the case for the plots in this thesis that used the elbow method. A cluster more or less would definitely also be an acceptable answer. Furthermore, a different amount of clusters might not have changed the performance much, based on the deviance plots.

### 5.3.3 Categorical clustering and entity embedding

Due to the random nature of NN the location of the features in CV 1 and CV 2 changed a lot between runs. This also affected the closeness of features. Taking the average location of ten trained NN can help keep the closeness of some features. However, some features might have a similar average location without them being close in any of the 10 NN models. This means non-similar features can be clustered together.

While two different sizes of hidden layers are tested and the input method for the NN, no grid search was performed to find the best parameters for the NN. This means that there possibly could have been a better NN configuration for this thesis.

Another issue with the trained NN is that its prediction of the claim frequency has a higher error than that of the GLM models. This shows that it is a worse model. Furthermore, the NN also produces negative claim frequencies, these are values that can not exist. As the entity embedding is trained with these NN models the closeness of features, found after training, might also be less accurate than desired or wrong, just like the predictions of the NN.

### 5.3.4 Pure premium

As mentioned in Chapter 2, this thesis only focuses on the pure premium. Pure premiums are used as an accurate estimation of the insurance company's risk for the given product. We have to note that the final pricing model used by the insurance company can contain a slightly different set of risk factors due to internal practices. Moreover, the pure premium, or the model based on it, is only a part of the full pricing of the insurance offered. The final price also contains additional considerations such as overheads, regulatory requirements, required profit margins and more, all of which are beyond the scope of this thesis.

Furthermore, the insurance company uses different software than the one used in this thesis when making their GLMs. This different software probably uses different parameters when fitting a GLM. This difference also means that the base model in this thesis, using the insurance companies clusters, can also be different from their actual model. Thus, the base model is only an estimation of the actual pure premium model of the insurance company. This is likely further influenced by the chosen split of training and testing data.

### 5.3.5 Coefficients distribution

While the cumulative presence plots of the model's coefficient sometimes show that the two models follow the same path, this could give a false sense of belief that the two models behave similarly for a given risk factor. Similar coefficients might represent the high end of a numerical risk factor in one model and the low end in the other.

This can also happen within a given model. Two neighbouring coefficients might represent very different sectors of a numeric risk factor. This can reduce the explainability of the pure premium to stakeholders, regulatory bodies and customers.

A better representation for future projects could be to show the different coefficient values for each of the available options in a variable. An example would be comparing coefficients of each individual feature in the categorical variables. Presenting this data in a clear way might be difficult with a large amount of features.

### 5.3.6 Average Coefficients

Looking at the average coefficients of the two models risk factors for the policies with the biggest price gives a general overview of what risk factors contribute to the price change. However, certain clusters in a given risk factor could have a larger effect. This nuanced information can be lost when using averages.

### 5.3.7 ML models

It is suggested to use continuous inputs when possible for the used ML models. While we wanted to know if the insurance company could keep the desired classes (for explainability reasons), this could have affected the performance of the models.

The low number of claims, 1%, could have impacted the performance of the models. These low amount of claims could also be a reason to use other ML models. Low occurrence is also present in the banking world when they need to predict the probability of default, the types of models the banking world uses might have been more suitable for this data set.

## 5.4 Implications for the insurance company

Based on the results of this thesis, we can give a set of recommendations to the Dutch insurance company and the insurance industry in general. At first, a recommendation is given on the use of data-driven clustering via entity embedding.

As the difference in MSE and MAE error is not present before the fourth decimal value, no model performs noticeably better than another. While a small difference is seen in the mean Poisson deviance, one of the insurance company's valuation metrics, this difference is not large enough for us to recommend a switch from expert clustering to data based clustering. This is due to the fact that for the categorical data based clustering, with entity embedding, the explainability that was present in the expert made clusters is lost. The entity embedding might be a helpful tool when the insurance company and the expert are unsure in what exact group a certain variable needs to be placed.

When looking at the data-driven clustering methods for the numerical variables, our recommendation for the insurance company would be to keep the expert's clustering method in their current model. This is due to the fact that the best fit method used to select the optimal amount of clusters does not always create a lower mean Poisson deviance.

However, the created clusters are still explainable and have the same MSE and MAE as the existing GLM. Due to this, the data-driven clustering method of either Fisher-Jenks or K-means can be used by insurance companies when making a new GLM. This could give the experts more time to work on the clustering of categorical variables.

Our final recommendation for insurance companies thinking of switching from a GLM to an ML model for their frequency prediction is to do more research into the subject. At the end of this thesis, we demonstrated that using the existing clusters as inputs into a standard ML models does not outperform the GLM. Thus, more research is needed on how ML methods could be implemented in an insurance company.

## 5.5 Future research

While we looked into multiple different aspects of data-driven clustering, several avenues for future research and further exploration are still available. In the following section, some of these avenues that appeared while during the thesis are highlighted.

We showed that two models with an almost identical performance on the overall data set can provide widely different estimates for individual policies. A certain trade-off in errors is expected at points where the model fits certain policies better and certain policies worse. This trade-off might also be present in the amount of clusters. A larger amount of clusters might lead to overfitting, while a smaller amount might have trouble capturing important relationships in the data. Future research could have a closer look into the reason for the large amount of change on individual level with no clear improvements in the overall model.

Even though the models seem almost identical, one could be a better model than the other at the policy level. Future research could look into this by back testing the models and comparing the models on new policy data. The difference in pure premium between these models for specific customer groups could showcase if one of these models performs better.

The data set used in the thesis contains a lot of 0 claim policies. This large number might invalidate the Poisson distributional assumptions. Research could be done into the use of different models that might better deal with the large number of 0 in this data set. Sarul, 2015 and Zhang et al., 2022 show promising results for these different types of models.

Future research can also look into testing the methods of this thesis on different data sets. Our first recommendation would be a similar type of insurance data set with a higher claim rate, this might change the performance of the different clustering methods and the GLM. The data set could either be an internal one or a publicly available set, such as the one used in Henckaerts et al., 2018.

Avanzi et al., 2023 notes that high cardinal categorical variables are present as risk factors in many types of insurance. One of the examples given was health insurance, with risk factors including the cause of injury to workers or their occupation. The insurance company offers a lot of different types of insurances and also has health insurance data sets. Research could be done into how well entity embedding and data-driven clustering work on such a different data set. As the health insurance data set is probably also larger, it would allow for the testing of the clusters on the severity predictions as well as the frequency predictions.

# 6 Conclusion

This project aimed to identify whether data-driven clustering can improve the modelling of the pure premiums of an insurance company. The research questions for this thesis are listed below.

1. "To understand the extent to which clustering within variables improves the current modelling method, of using a GLM, for determining the pure premium. Furthermore, we look at how the GLM modelling technique compares to other more novel and future forward machine learning modelling techniques for the pure premium when using clustered data."

The sub research questions are as follows:

1. Which are the current methods to handle numerical and categorical data for GLMs?
2. How can clusters be made, along with determining their optimal number, to ensure that each cluster significantly influences the pure premium?
3. How do combinations and variations of numerical and categorical clustering impact the pure premium?
4. How does pure premium modelled by the GLM compare to novel modelling techniques?

The research questions are answered in the discussion in Chapter 5. It can be concluded from the previous chapter that the data-driven clustering techniques can improve the GLM for predicting the frequency of insurance claims when looking at the chosen evaluation metric of the Poisson deviance. However, this result was only obtained by a combination of clusters that individually lowered the mean deviance and not based on how well the clusters explained the variance of the data. Furthermore, while a slightly better mean deviance was obtained, no difference was seen in the MSE and MAE, looking at the fourth decimal values, when compared to the existing model of the insurance company.

The improvements obtained in the mean Poisson deviance of the newer GLM model does come with some drawbacks. Switching to this newer model would cause a large change in the individual pure premium of policies without a clear increase in the model's performance. Based on this and the lower explainability of the newer GLM, we concluded that the insurance company should stay with the current model. We also reached this conclusion regarding whether the insurance company should adopt ML models using the clustered data as inputs. We found this conclusion because, with the clustered inputs, none of the tested ML models outperformed any of the GLMs.

While we did not find clear improvements on the current modelling technique of the pure premium, our findings can still contribute to the insurance industry. We showed that different clustering techniques can be used to match the results obtained by clusters made by experts. Furthermore, it highlighted the importance of an expert's judgment when clustering a categorical variable, as the data-driven technique does not take explainability into account. Ultimately, prioritizing the explainability of models within the insurance industry is of more importance than opting for a marginally better model at the cost of the explainability.

# 7 Appendix

## 7.1 Clustering results

This show the results of the entity embedding of the neural network with a hidden layer that each contain 120 neurons. The embedding space is reduced to a two dimensional space using t-SNE (van der Maaten and Hinton, 2008).

### 7.1.1 Categorical variable 2

The first two images are the results of the NN trained using continuous inputs along with the two categorical inputs. Of these the first image shows the cluster results found by the insurance company. The colour of each point corresponds with its cluster. In the second image the results of the K-mean clustering, for 27 clusters, is shown. It can be seen that the points in the plot belong to different clusters.

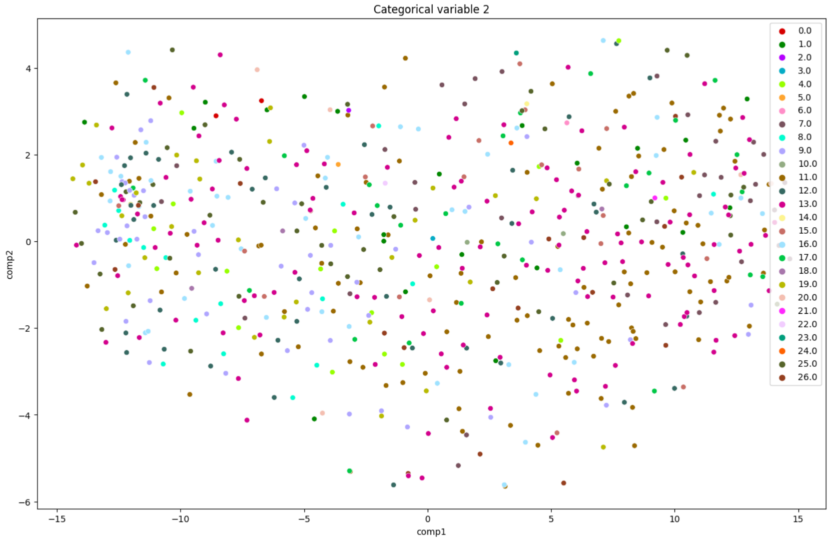


Figure 24: Location of variables in CV 2 in a reduced embedded space.
Clustered using the insurance companies clustering. Using continuous data as input for the NN

The second set of images show the result of the NN trained using all inputs as embedded inputs (based on the insurance companies existing grouping). Here the first images also shows the cluster results for CV 2 found by the insurance company. The second image shows the results found using the K-means cluster method based on the location of the points in the embedded space (pre-reduction due to the t-SNE method).

## 7.2 Mean Poisson deviance of GLM with CV 1 and CV 2

The first two figures of the section show the Mean Poisson deviance of GLM with cluster found using entity embedding. These where found by the trained NN using continuous data and a hidden layer size of 120. The first figure shows the results of CV 1 and the second that of CV 2.

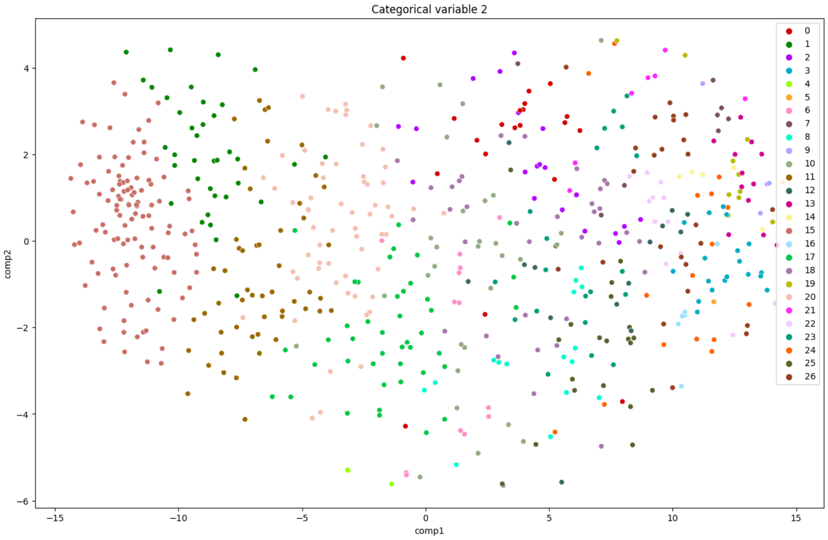


Figure 25: Location of variables in CV 2 in a reduced embedded space.
Clustered using the K-means clustering method. Using continuous data as input for the NN

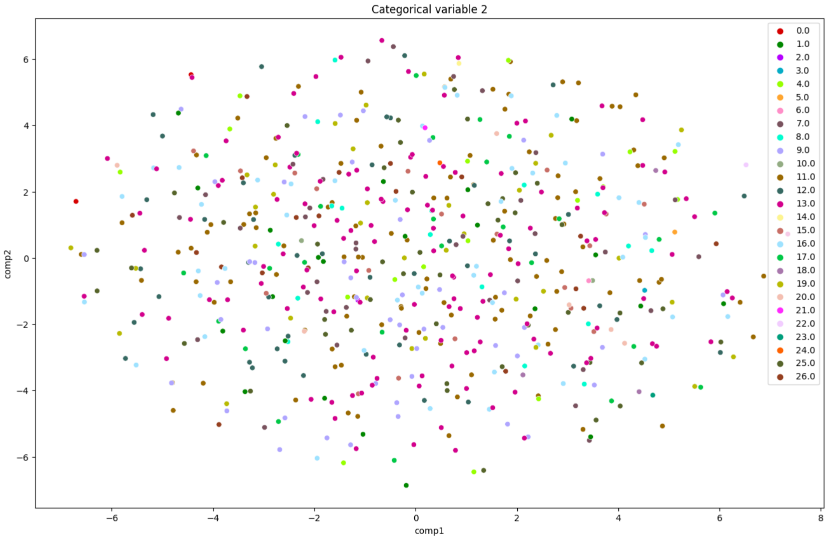
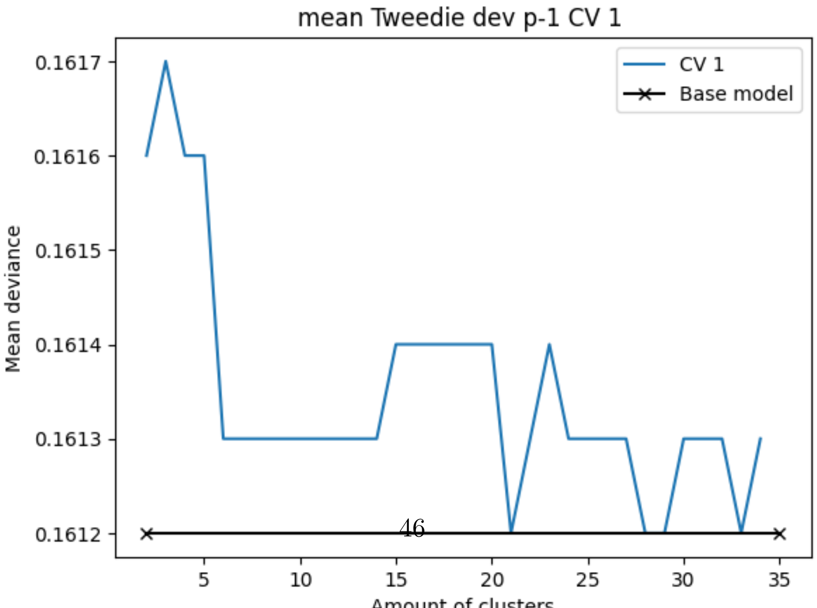


Figure 26: Location of variables in CV 2 in a reduced embedded space.
Clustered using the insurance companies clustering. Using categorical data as input for the NN



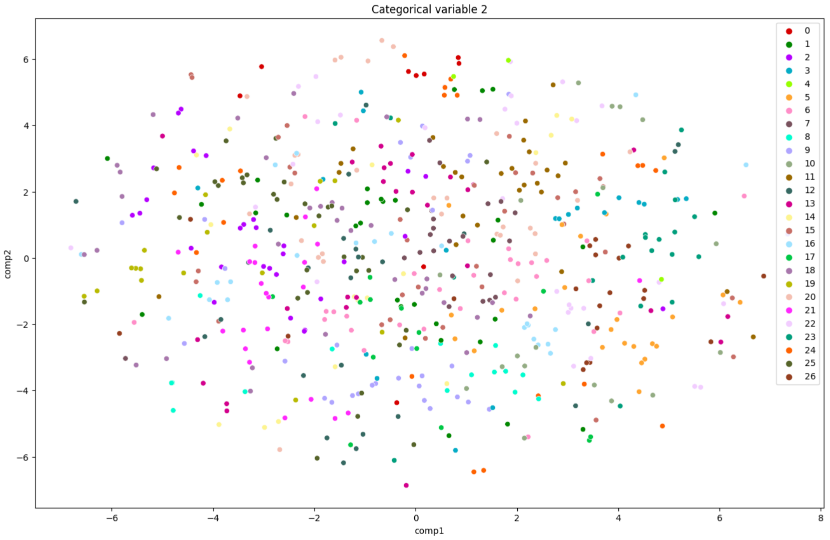


Figure 27: Location of variables in CV 2 in a reduced embedded space.
Clustered using the K-means clustering method. Using categorical data as input for the NN

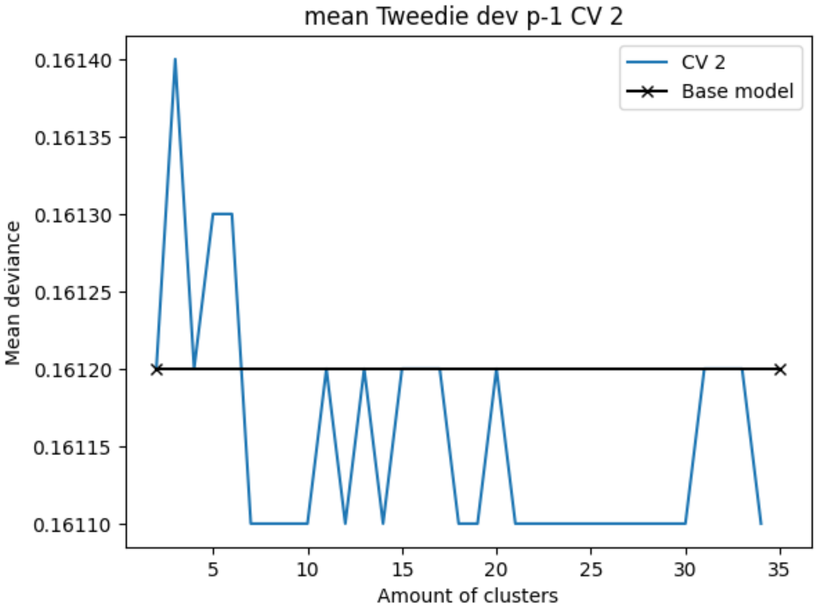


Figure 29: Mean Poisson deviance of the GLM with CV 2 clusters. NN with hidden layers of 120.

The following figures show the Mean Poisson deviance of GLM with cluster found using entity embedding. These where found by the trained NN using categorical data. The first figure shows the results of CV 1 and the second that of CV 2.

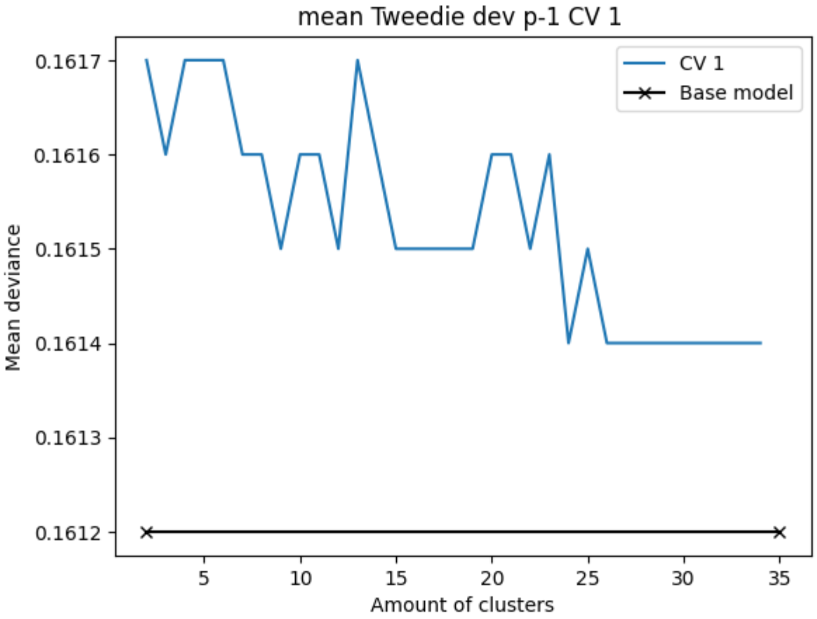


Figure 30: Mean Poisson deviance of the GLM with CV 1 clusters. NN trained on categorical data.

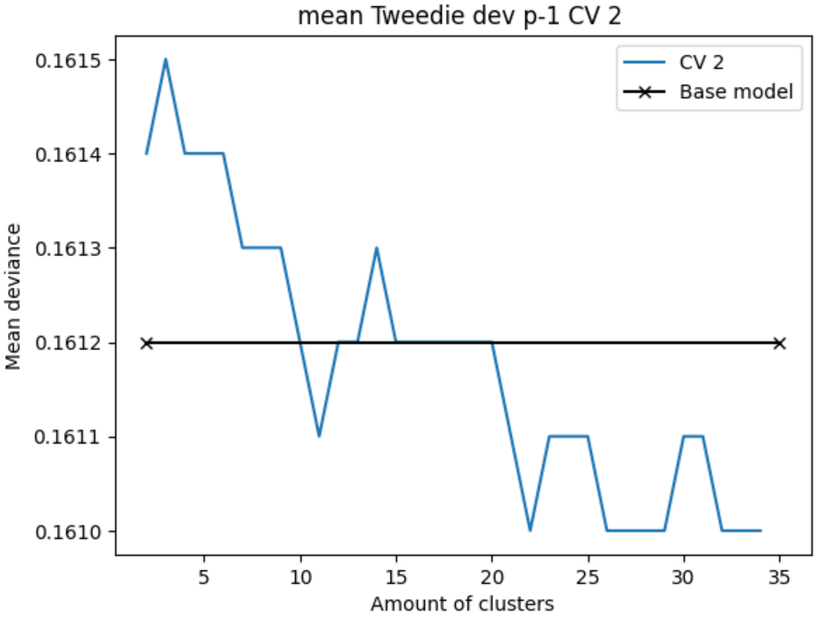


Figure 31: NN trained on categorical data.

### 7.3 Cumulative distribution of the weights in the GLM

The cumulative distribution of the remaing weights of the base GLM and the picked model.

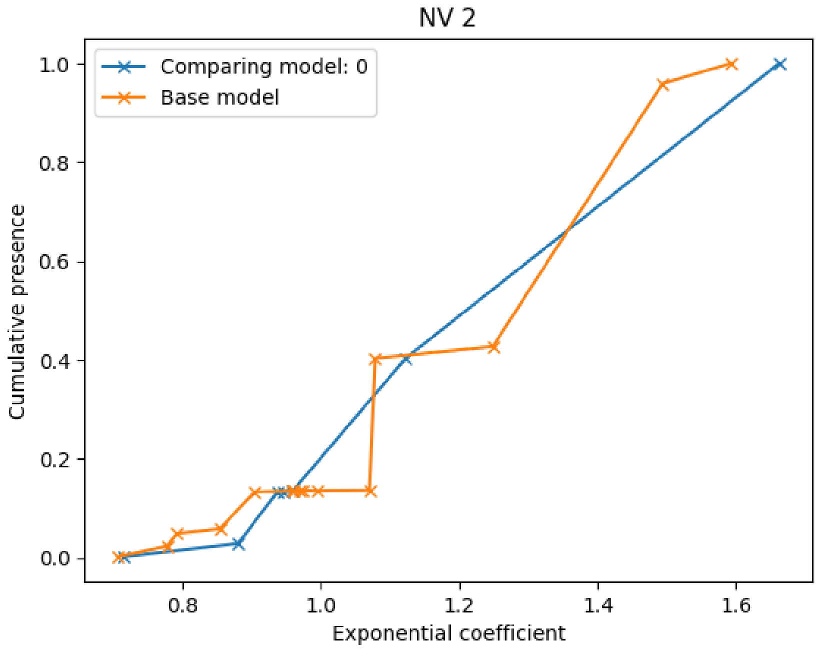


Figure 32: Cumulative distribution of the weights of NV 2 for the GLMs.

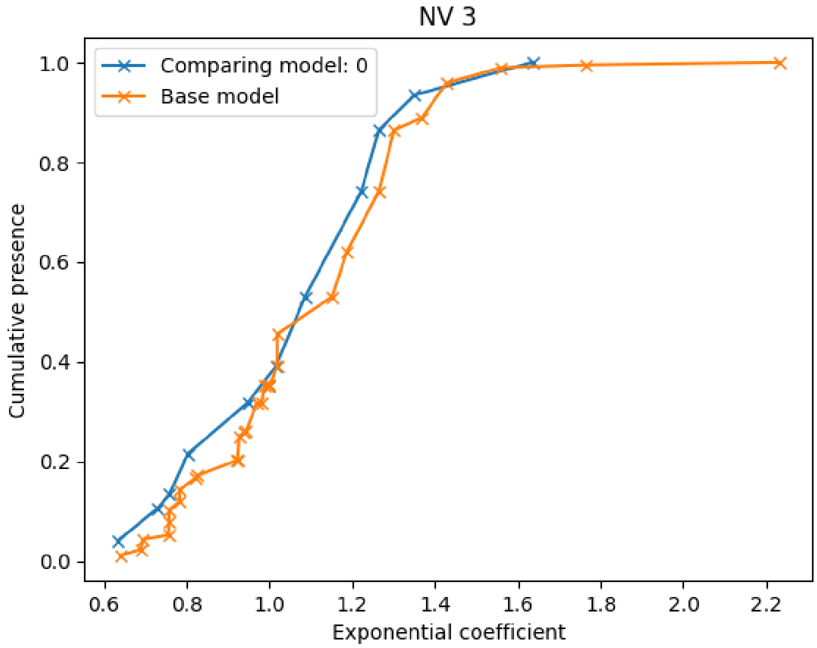


Figure 33: Cumulative distribution of the weights of NV 3 for the GLMs.

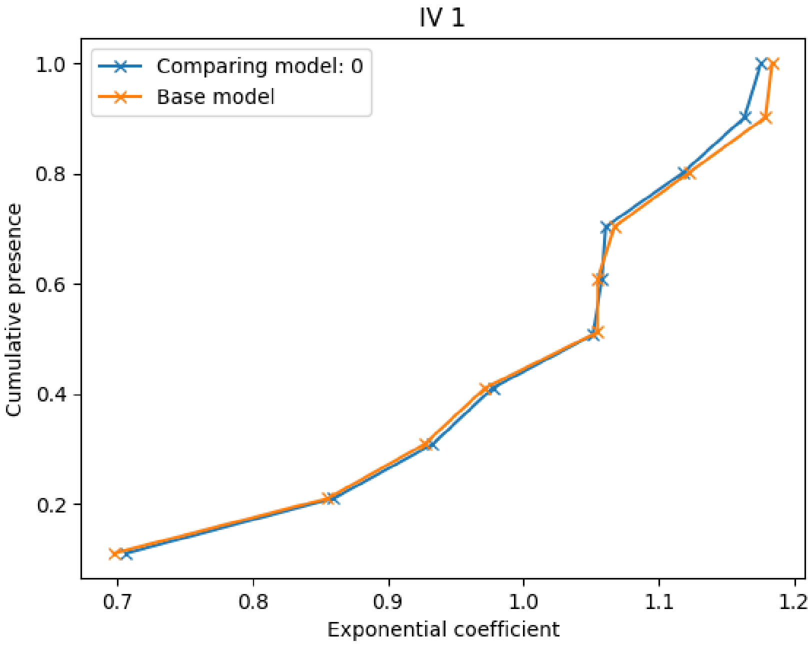


Figure 34: Cumulative distribution of the weights of IV 1 for the GLMs.

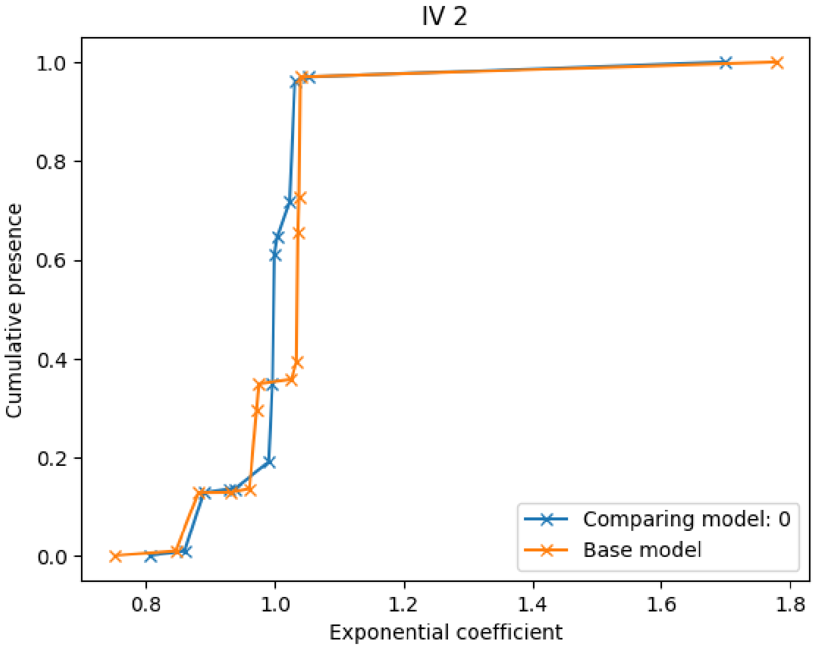


Figure 35: Cumulative distribution of the weights of IV 2 for the GLMs.

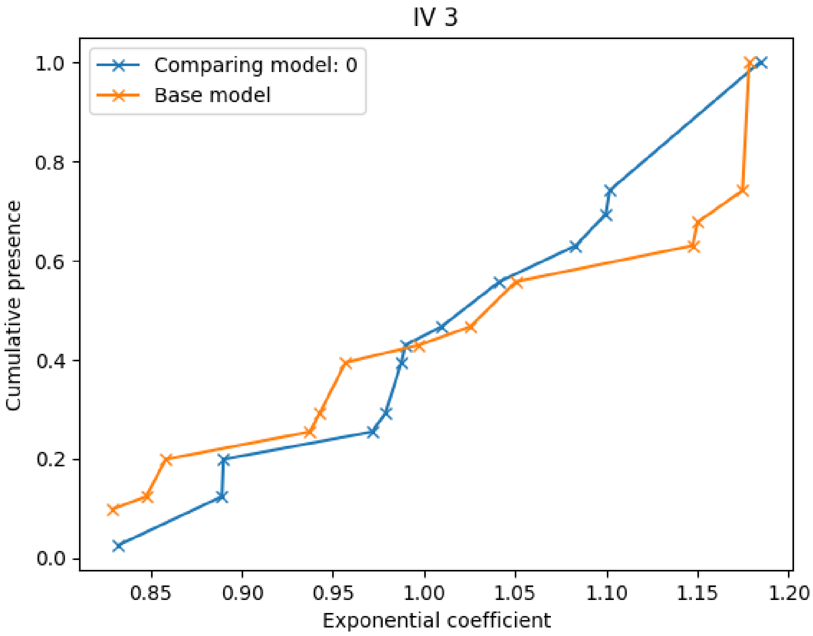


Figure 36: Cumulative distribution of the weights of IV 3 for the GLMs.

# References

AFM & DNB. (2019). Aandachtspunten AFM en DNB bij artificiële intelligentie in de verzekeringssector. https://www.afm.nl/nl-nl/sector/actueel/2019/jul/verkenning-ai-verzekeringssector

Anderberg, M. R. (2014). *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks* (Vol. 19). Academic press.

Avanzi, B., Taylor, G., Wang, M., & Wong, B. (2023). Machine learning with high-cardinality categorical features in actuarial applications. https://doi.org/10.48550/arXiv.2301.12710

Blazenko, G. (1986). The economics of reinsurance. *Journal of Risk and Insurance*, 258–277.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010950718922

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Charters de Azevedo, F., Oliveira, T., & A., O. (2016). Modeling non-life insurance price for risk without historical information. *Revstat - Statistical Journal*, *14*, 171–192. https://doi.org/10.57805/revstat.v14i2.185

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Coutts, S. M. (1985). Risk theory—the stochastic basis of insurance. by r. e. beard, t. pentkainen amp; e. pesonen. published by chapman amp; hall. *Journal of the Institute of Actuaries*, *112*(2), 303–304. https://doi.org/10.1017/S0020268100042153

de Bont, D. (2022, December). Geographical risk in the dutch car insurance : A data-driven approach to measure regional effects on the claim frequency. http://essay.utwente.nl/93997/

de Jong, P., & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755408.002

Denuit, M., & Marechal, X. (2007, November). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. https://doi.org/10.1002/9780470517420

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap* (1st). Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, *53*, 789–798. https://api.semanticscholar.org/CorpusID:120138631

Frees, E., & Valdez, E. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, *103*, 1457–1469. https://doi.org/10.1198/016214508000000823

Frees, E. W. (2014). Frequency and severity models. In E. W. Frees, R. A. Derrig, & G. Meyers (Eds.), *Predictive Modeling Applications in Actuarial Science* (pp. 138–164, Vol. 1). Cambridge University Press. https://doi.org/10.1017/CBO9781139342674.006

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/AOS/1013203451

Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods.

Gao, G., Meng, S., & Wüthrich, M. (2018). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, *2019*, 1–20. https://doi.org/10.1080/03461238.2018.1523068

Glorot, X., Bordes, A., & Bengio, Y. (2010). Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, *15*. https://www.researchgate.net/publication/215616967_Deep_Sparse_Rectifier_Neural_Networks

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. https://doi.org/https://doi.org/10.48550/arXiv.1604.06737

Haberman, S., & Renshaw, A. (1996). Generalized linear models and actuarial science. *The Statistician*, *45*, 407. https://doi.org/10.2307/2988543

Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681–705. https://doi.org/10.1080/03461238.2018.1429300

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, *25*(2), 255–285. https://doi.org/10.1080/10920277.2020.1745656

IBM. (2023). *Decision trees* [Accessed on Date of access]. IBM. https://www.ibm.com/topics/decision-trees

Jenks, G. F. (1967). The data model concept in statistical mapping. https://api.semanticscholar.org/CorpusID:215850874

Jørgensen, B., & Paes De Souza, M. C. (1994). Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, *1994*(1), 69–93.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008, January). *Modern Actuarial Risk Theory: Using R*. https://doi.org/10.1007/978-3-540-70998-5

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441–458. Retrieved October 18, 2023, from http://www.jstor.org/stable/2486927

Kilroy, A. (2022, May). 8 different types of insurance policies and coverage you need. https://www.forbes.com/advisor/insurance/types-of-insurance-policies/

Kirkpatrick, E. (2022, January). Heidi Klum says her legs are insured for $2 million, but one is "more expensive" than the other. https://www.vanityfair.com/style/2022/01/heidi-klum-legs-insured-2-million-breasts-ellen-degeneres

Kuhn, M., & Johnson, K. (2013, January). *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Liedtke, P. (2007). What's insurance to a modern economy? *The Geneva Papers on Risk and Insurance - Issues and Practice*, *32*, 211–221. https://doi.org/10.1057/palgrave.gpp.2510128

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Louppe, G. (2015). Understanding random forests: From theory to practice.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. https://api.semanticscholar.org/CorpusID:6278891

Masci, P. (2011). The history of insurance: Risk, uncertainty and entrepreneurship. *Business and Public Administration Studies*, *6*(1), 25–25.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor. Newsl.*, *3*(1), 27–32. https://doi.org/10.1145/507533.507538

Ostaszewski, K. (2018). The social purpose of insurance and why it matters. *HAWAII UNIVERSITY INTERNATIONAL CONFERENCES*. https://huichawaii.org/wp-content/uploads/2018/01/Ostaszewski-Krzysztof-2018-AHSE-HUIC.pdf

Park, H., Borde, S. F., & Choi, Y. (2002). Determinants of insurance pervasiveness: A cross-national analysis. *International Business Review*, *11*, 79–96. https://doi.org/10.1016/S0969-5931(01)00048-8

Park, S. C., & Lemaire, J. (2012). The impact of culture on the demand for non-life insurance. *Astin Bulletin*, *42*, 501–527. https://api.semanticscholar.org/CorpusID:56388983

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://api.semanticscholar.org/CorpusID:13252401

Rijksoverheid. (2023, February). Verplichte zorgverzekering afsluiten. https://www.rijksoverheid.nl/wetten-en-regelingen/productbeschrijvingen/verzekeringsplicht-zorgverzekering

Robbins, H. E. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407. https://api.semanticscholar.org/CorpusID:16945044

Rodríguez, P., Bautista, M. A., Gonzàlez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, *75*, 21–31. https://doi.org/https://doi.org/10.1016/j.imavis.2018.04.004

Ross, S. (2022). What is adverse selection in the insurance industry? https://www.investopedia.com/articles/personal-finance/080616/what-adverse-selection-insurance-industry.asp

Sarul, S. (2015). An application of claim frequency data using zero inflated and hurdle models in general insurance. *Pressacademia*, *4*, 732–732. https://doi.org/10.17261/Pressacademia.2015414539

Schlesinger, H. (2013). The theory of insurance demand. In G. Dionne (Ed.), *Handbook of insurance* (pp. 167–184). Springer New York. https://doi.org/10.1007/978-1-4614-0155-1_7

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

ScikitLearn. (2023a). 1.1. Linear Models. https://scikit-learn.org/stable/modules/linear_model.html#solvers

ScikitLearn. (2023b). 1.10. Decision Trees. https://scikit-learn.org/stable/modules/tree.html

Shi, P., & Shi, K. (2022). Non-life insurance risk classification using categorical embedding. *North American Actuarial Journal*, *0*(0), 1–23. https://doi.org/10.1080/10920277.2022.2123361

Thorndike, R. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Trenerry, C. F. (1926). The origin and early history of insurance including the contract of bottomry. *Journal of the Institute of Actuaries*, *57*(2), 277–278. https://doi.org/10.1017/S0020268100031176

Upton, G., & Cook, I. (2008). *A Dictionary of Statistics*. Oxford University Press. https://doi.org/10.1093/acref/9780199541454.001.0001

van der Maaten, L., & Hinton, G. (2008). Viualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wikipedia. (2023). *Artificial neural network* [Accessed on Date of access]. Wikipedia. https://en.wikipedia.org/wiki/Artificial_neural_network

Wuthrich, M. V., & Buser, C. (2016). *Data analytics for non-life insurance pricing* (Swiss Finance Institute Research Paper Series No. 16-68). Swiss Finance Institute. https://EconPapers.repec.org/RePEc:chf:rpseri:rp1668

Wüthrich, M. V., & Merz, M. (2019). EDITORIAL: YES, WE CANN! *ASTIN Bulletin*, *49*(1), 1–3. https://doi.org/10.1017/asb.2018.42

Yang, Y., Qian, W., & Zou, H. (2016). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models.

Zhang, P., Pitt, D., & Wu, X. (2022). A comparative analysis of several multivariate zero-inflated and zero-modified models with applications in insurance.