EUROPEAN CONSORTIUM FOR MATHEMATICS IN INDUSTRY

ECMI Modeling Week 2023
University of Szeged

CAR INSURANCE PREMIUM CALCULATION

Project Team:

Arianna Zanin	(University of Milan)
Koppány Dodony	(University of Szeged)
Filip Rydin	(Chalmers University of Technology)
Roeland Hooijmans	(Eindhoven University of Technology)
Tiago Miranda	(University of Lisbon)

Instructor:
Dora Selesi (University of Novi Sad)

August 1, 2023

Contents

1 Introduction

Actuarial mathematics is widely applied in the insurance sector. The purpose is to employ a data-driven approach to, among other tasks, calculate fair premiums for different policy holders. This allows the insurance company to distinguish high risk from low risk individuals, and in turn give more competitive premiums to customers less likely to file claims. More in general, knowing which variables and characteristics influence each other and the expected claim amount can give great insights about customers that can be used for a number of tasks and improve business performance.

One particular challenge when dealing with insurance data and the insurance industry is that there are both ethical and legal constraints on the methods used to calculate premiums (2). Results must be explainable and statistically sound, so that a given policy holder understands why they pay a certain premium. Moreover, to prevent discrimination, certain characteristics should not be used to separate individuals. Such example are gender and ethnicity, which are forbidden to use in certain countries.

In this report, our objective is to develop a comprehensive methodology for premium calculation in the context of claim data analysis. We begin by calculating a manual premium based on the collective claim data of the entire population. We explore various statistical models and tests, such as ANOVA, Tukey tests, chi-squared tests, and correlation tests (such as Pearson's correlation coefficient) to determine the risk classes and define the factors that should be considered and those that should be excluded. These analyses help us identify the key features that significantly influence the total claim amount. Moving forward, in the next step, we investigate how to incorporate an individual's personal history into their premium calculation. This involves incorporating a credibility factor that accounts for an individual's unique claims history. By employing these statistical techniques and incorporating individual history, our approach aims to provide a robust and personalized premium calculation framework.

2 Theoretical model

We model the total loss for a given policyholder in a given year as

$$S = X_1 + X_2 + \cdots + X_N \tag{1}$$

where N is a discrete random variable and $X_1, \cdots, X_N$ are i.i.d. non-negative continuous random variables. X_i are claim sizes and N is the number of claims. Making use of the tower rule, it's easy to show that the expected value of S, that is the the manual premium M we want to compute, can be expressed in this form:

$$M = \mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S|N]] = \cdots = \mathbb{E}[N]\mathbb{E}[X]. \tag{2}$$

In the same way we can express the variance as:

$$Var(S) = \mathbb{E}[N]\ Var(X) + Var(N)\ (\mathbb{E}[X])^2. \tag{3}$$

In addition, an explicit relation between the specific distributions of S, N and X can be given in terms of moment generating functions:

$$G_S(t) = E[t^S] = E[t^{\sum_{i=0}^{N} X_i}] \tag{4}$$

$$= \sum_{n=0}^{\infty} E[t^{\sum_{i=0}^{N} X_i} | N = n] P(N = n) \tag{5}$$

$$= \sum_{n=0}^{\infty} (G_X(t))^n \cdot P[N = n] = \tag{6}$$

$$= G_N(t) \circ G_X(t). \tag{7}$$

Here $G_N(\cdot)$ is the probability generating function of the discrete variable N.

3 Methods

In this section the overall methodology of all steps in the modelling is discussed. Since each step builds on the previous ones, the result of all the steps were allowed to impact the initial plan to a great extent.

3.1 Manual premium calculation

In this part of the modelling, the population was considered as a whole. The individual distributions and means for number of claims, N, and claim size, X, were estimated. This, in turn, allowed us to estimate a global premium and the distribution of the total yearly claim amount.

3.1.1 Claim size distribution

For modeling claim sizes in car insurance, several distributions are commonly used to capture the characteristics of the data. A key characteristic that often emerges in claim size data is the presence of heavy-tailed distributions, which signify a higher likelihood of extreme or large losses compared to distributions with lighter tails. We investigated a number of different distributions: Gamma, Pareto, Exponential, F-distribution (centered and non-centered), χ^2 (centered and non-centered), Beta, Log-normal, Weibull, Gumbel, Logistic and Burr.

We compared the adequacy of the distributions by creating QQ-plots, histogram vs pdf plots and calculating the Bayesian Information Criterion:

$$\text{BIC} = k \log(n) - 2 \log(\hat{L}), \tag{8}$$

where k is the number of parameters in the model, n is the number of data points and $\hat{L}$ is the maximized likelihood of the data. We also tried to use a Kolmogorov-Smirnov test, which rejected every attempted distribution. It was determined that this was due to the exceptionally large sample size, which caused even small deviations to become significant. The Kolmogorov-Smirnov test was therefore not used in the end.

3.1.2 Number of claims distribution

For modeling the counting variables (in this case, counting the number of claims) several distributions are commonly used. We tested several distributions to find their parameters that best fit the data. We tested the following distributions: Poisson, Geometric and Negative Binomial.

As before we compared the adequacy of distributions using histograms and by calculating the Bayesian Information Criterion (BIC).

3.2 Risk class clustering

The objective of the second task was to identify significant risk profiles within the population by identifying relevant risk predictors. In pursuit of this goal, we conducted the following statistical instruments to filter out irrelevant factors (1).

- Pearson's correlation coefficient
- One-way ANOVA
- Two-way ANOVA
- Tukey's test
- Contingency tables with χ^2 tests

The predictors available in the data were the numerical variables income, number of family members and vehicle mileage, as well as the categorical variables sex, territory, driving experience, education, vehicle color, vehicle production year and vehicle manufacturer. In order to better compare numerical and categorical variables, we discretized the first into classes.

To have a better insight into the number of meaningful partitions available in the data we used t-SNE plots with different perplexity values. This is done to visualise high dimensional data in lower dimensions. The perplexity value is the effective number of neighbours considered when building the adjacency graphs between points. Therefore, when using a large value of perplexity the data tends to conglomerate in a single cluster (as datapoints are connected to growing number of neighbours) and as the value used decreases it separates into several clusters. The value of this parameter is purely up to

the researcher but analyzing how the plot changes with different perplexity values is a visual aid to understand the number of clusters present in the dataset.

3.3 Claim history modeling

In the third modeling step we also took personal claim history into account. This was done through two similar but separate methods, the European credibility method and the American credibility method. In both methods we assume that we know the amount of claims, N, a person has filed during k years.

3.3.1 European credibility method

The European credibility method is based on Bayesian statistics. We obtain a posterior distribution of a driver in the previously defined risk classes given its claim history. The premium calculation is based on the posterior distribution obtained. The risk premium is defined as

$$\text{Premium} = \mathbb{E}_\Lambda \left[\mathbb{E}[S|\Lambda] \mid N_1, N_2, \cdots, N_k \right]. \tag{9}$$

This expectation can be calculated using the total law of expectation as

$$\mathbb{E}_\Lambda \left[\mathbb{E}[S|\Lambda] \mid N_1, N_2, \cdots, N_k \right] = \sum_i \mathbb{E}[S|\Lambda = \lambda_i] \, \mathbb{P}(\Lambda = \lambda_i | N_1, N_2, \cdots, N_k) \tag{10}$$

where $\Lambda = \lambda_j$ means that the policyholder belongs to the risk class j.

Assuming independence of observations N_i the posterior distribution can be calculated by Bayes' theorem as follows

$$\mathbb{P}\left[\Lambda \mid N_1, N_2, \cdots, N_k \right] = \frac{\prod_{i=1}^{k} \mathbb{P}(N_i \mid \Lambda) \, \mathbb{P}(\Lambda)}{\mathbb{P}(N)}. \tag{11}$$

The likelihood, $\mathbb{P}(N_i \mid \Lambda)$, is given by the Poisson distribution with parameter λ_j depending on risk class, j. The prior, $\mathbb{P}(\Lambda)$, was chosen as the uninformative prior. The denominator is simply a constant, which can be initially ignored. The expression for the posterior distribution is, as such, simple to calculate given a known policy holder history.

After the posterior distribution has been calculated, the premium can be calculated, since $\mathbb{E}[S|\Lambda = \lambda_j]$ for all j is known from the second modelling step.

3.3.2 American credibility method

The American credibility method requires to fix two parameters p and r in order to find a credibility factor $\mathcal{Z}$ such that

$$\mathbb{P}(|\text{Premium} - M| \leq rM) \geq p, \tag{12}$$

where

$$\text{Premium} = \mathcal{Z}\bar{S}_k + (1-\mathcal{Z})M. \tag{13}$$

The variable $\bar{S}_k$ encodes the past history of the client, specifically

$$\bar{S}_k = \frac{\sum_{i=1}^{k} S_i}{k} \tag{14}$$

where S_i is the recorded total loss for the specific policyholder in the i-th year. We derived the following explicit formula for $\mathcal{Z}$ if the numbers of years considered is k:

$$\mathcal{Z} = \frac{-\sqrt{k}rM}{\sqrt{Var(S)}\ \Phi^{-1}\left(\frac{1-p}{2}\right)}. \tag{15}$$

Here M is the manual premium calculated in the previous step for the risk class of the policyholder. A rigorous proof of 15 is reported in the following subsection.

It is crucial to highlight that in order to derive the formula we assumed the number of years k to be large enough to apply the CLT (Central Limit Theorem) and consider $\bar{S}_k$ as normally distributed. It is therefore important to notice that what we obtained is an approximation valid only for histories that are large enough.

3.3.3 Proof of the American credibility factor formula

Let's start noticing that $\mathbb{P}(|\mathcal{Z}\bar{S}_k + (1-\mathcal{Z})M - M| \leq rM) \geq p$ if and only if

$$\mathbb{P}(|\mathcal{Z}(\bar{S}_k - M)| > rM) < 1-p. \tag{16}$$

This is also equivalent to requiring

$$\mathbb{P}(\mathcal{Z}(\bar{S}_k - M) > rM) + \mathbb{P}(\mathcal{Z}(\bar{S}_k - M) < -rM) < 1-p, \tag{17}$$

that is

$$\mathbb{P}\left(\bar{S}_k > \frac{rM}{\mathcal{Z}} + M\right) + \mathbb{P}\left(\bar{S}_k < \frac{-rM}{\mathcal{Z}} + M\right) < 1-p. \tag{18}$$

We can suppose that k is large enough in order to apply the Central Limit Theorem and suppose that $\bar{S}_k$ is normally distributed. After the normalization step, because of the fact that in our setting $E[\bar{S}_k] = M$, we obtain

$$\mathbb{P}\left(\mathcal{N}(0,1) > \frac{rM}{\sigma\mathcal{Z}}\right) + \mathbb{P}\left(\mathcal{N}(0,1) < -\frac{rM}{\sigma\mathcal{Z}}\right) < 1-p. \tag{19}$$

This leads us to the inequality

$$\Phi\left(-\frac{rM}{\sigma\mathcal{Z}}\right) < \frac{1-p}{2} \tag{20}$$

that is finally solved by

$$\mathcal{Z} > -\frac{rM}{\sigma\Phi^{-1}\left(\frac{1-p}{2}\right)}. \tag{21}$$

Notice that $\sigma = \frac{\sqrt{Var(S)}}{\sqrt{k}}$ so that we obtained the formula.

4 Results

4.1 Manual premium calculation

Based on the BIC and inspection of the plots, we found the Burr-distribution to be the best fit for the data, followed by a non-centered F-distribution. The pdf vs histogram plot and the QQ-plot of the Burr distribution can be seen in Figure 1. As can be seen from the QQ-plot, the Burr distribution has a heavier tail than the actual data. From a risk-management standpoint, this might be beneficial, since the risk of large claim sizes is overestimated compared to reality.

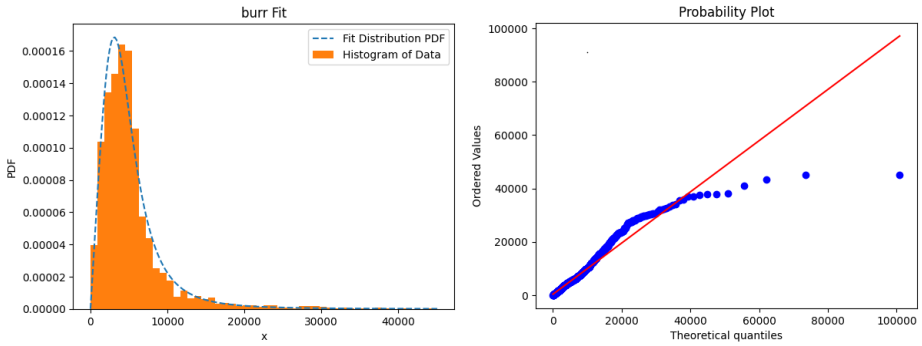

Figure 1: The estimated Burr distribution pdf compared to the histogram of the data and the quantile to quantile plot.

The estimated parameters for the Burr distribution are

$$\hat{c} \approx 2.802 \quad \hat{d} \approx 0.723 \quad \text{loc} \approx 4.642 \quad \text{scale} \approx 4930.$$

For the number of claims (N), based on the BIC metric and visual inspection of the plots, we concluded that all 3 options were viable. After taking a closer look into the literature available we chose the Poisson distribution as it's widely used in this field and documentation is plentiful. The estimated parameter is $\hat{\lambda} \approx 0.04521$. For the pmf vs histogram plot, see Figure 2.

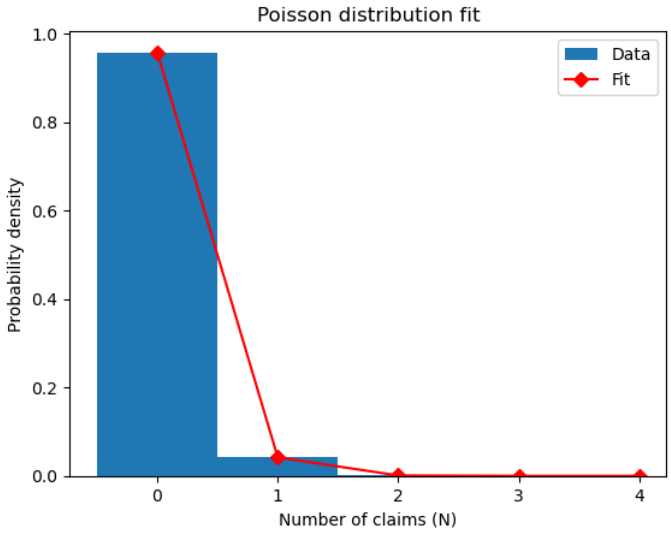

Figure 2: The estimated Poisson distribution pmf compared to the histogram of the data.

The mean of the number of claims in the data was calculated to 0.04521, whereas the mean of the claim size was calculated to 5322.27 euros. According to (2), a suitable manual premium is then 240.6 euros.

Based on the result of the distribution for claim size and the number of claims, the moment generating function for the total claim amount can be calculated as a composition according to (4). This uniquely defines both the distribution and the pdf. As such, the final distribution for the total claim amount was determined to be Poisson-Burr.

4.2 Risk class clustering

To classify individuals in different risk profiles we needed to assess the independence between each variable and a response variable. In this scenario the response variable is the total claim amount (S). However, we opted to subdivide the problem at hand by analysing the relationship between variables and the per-user average claim amount and between variables and the number of claims. Then, if the average claim amount is independent from a certain variable, the analysis of independence between that variable and the number of claims transfers to the total amount of claims.

We based our approach of identifying the key variables to classify customers on the statistical tests previously mentioned and on data visualizing algorithms. Here we show the most relevant results found.

The results of the correlation analysis between each explanatory variable and the average claim size is in Table 6 in the Appendix. These results were obtained after performing one-hot encoding of the categorical variables and we found no evidence that any variable was strongly correlated to the average claim size. Therefore, from

this point on wards, we used the study of independence from the number of claims as a proxy for the study of the independence from the total claim amount.

Feature	**Number of claims**
Mileage	$2.2 \cdot 10^{-16}$
Manufacturer	0.02241
Color	0.0005902
Production Year	0.2538
No. of Family Members	0.9035
Income	0.02482
Education	0.01252
Driving Experience	$2.2 \cdot 10^{-16}$
Territory	0.0614
Sex	0.1027

Table 1: p-values of χ^2 test for independence

From Table 1 we can conclude that after performing univariate analysis the candidate variables for the clustering are: Mileage, Color, Income, Education and Driving Experience.

However, when combining variables and retesting, these relationships do not hold. We concluded that Driving Experience was the only relevant information needed to divide the population into different risk profiles and adding extra information was not necessarily beneficial. Moreover, we concluded that the difference between senior, business and experienced drivers was not significant. As such, these drivers were bundled together in one risk class.

In order to have a better insight into the number of meaningful partitions available in the data we used t-SNE plots. Running the algorithm with perplexity = 5 (5) seemed to give the best output. Luckily our trials with other perplexity values had not led us to mentionably different outcomes. This method is not as precise as running statistical tests, but from different plots we can conclude which variables are worth testing for correlation and independence, since the amount of clusters (what are may not be strongly separable) are corresponding to variables that have some kind of similarities. We have made 2D to 2D plots, so we are not emphasizing the dimension reduction part, we are rather looking for similarities(6) in the data.

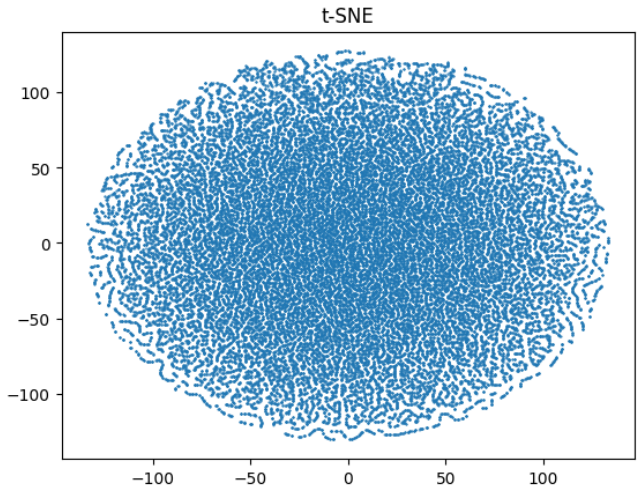

Figure 3: Output of t-SNE for the whole database

From the output in Figure 3 we can derive that there are no easily identifiable clusters in terms of the whole data. After this, we tested the relation between the number of claims and different categorical variables to identify risk classes. We observe the driving experience in Figure 4.

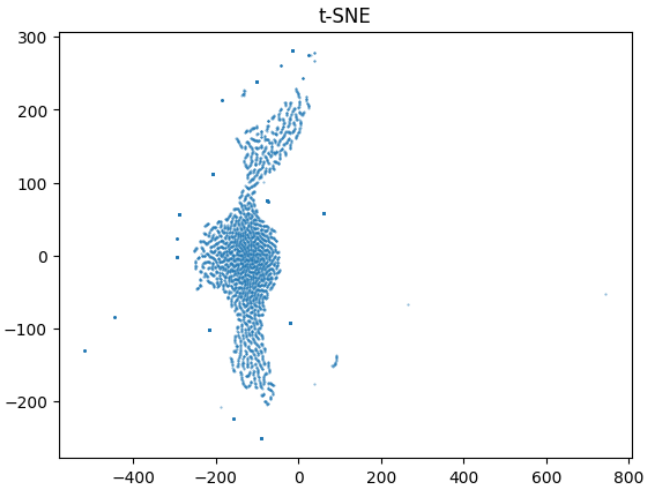

Figure 4: Output of t-SNE for driving experience and the number of claims

We can come to the conclusion, that making 3 risk classes based on experience is a reasonable decision.

The next figure shows another result in terms of mileage. From this we can see that making another 2 risk classes may not be sufficient. We can strongly separate 2 sets with a line, the distances between those sets are much lower than the diameter of the sets though.

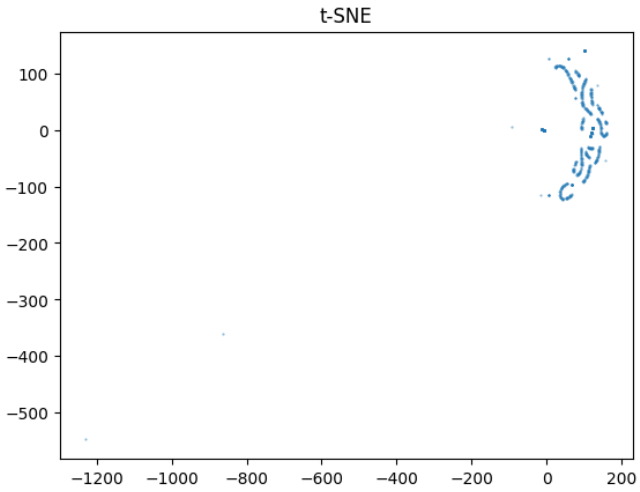

Figure 5: Output of t-SNE for mileage and the number of claims

When testing color, education, manufacturer, sex and territory respectively, the results are similar. We did not find any similarities between the variables mentioned above and the number of claims. For the rest of the t-SNE plots see Figure 6-10 in the appendix.

The final risk classes, as well as key statistics for the different groups, can be seen in table 2. Note that the majority of drivers belong to the experienced risk class.

	Trial	Beginner	Experienced
$\mathbb{E}[N]$	0.1022	0.0695	0.0420
$\mathbb{E}[X]$	6008.76	5166.53	5271.46
$\mathbb{E}[S]$	614.28	358.82	221.29
Count	3453	4190	92357

Table 2: Key statistics of the three different risk classes.

4.3 Claim history modelling

The two methods to take into account a persons past history work slightly differently, since the American method takes into account the risk class of the insured and the European method does not. Nevertheless, the two methods were compared for three different cases of past history to analyze their advantages and disadvantages.

4.3.1 European credibility method

We used an uninformative prior with the incidence of each risk class in the population as the probabilities.

$$\mathbb{P}(\Lambda) = [0.03453,\ 0.04190,\ 0.92357] \tag{22}$$

For a given history $\mathcal{H} = [0, 0, 1, 0, 0, 0, 0, 0, 1, 0]$ for 10 years the posterior distribution is:

$$\mathbb{P}\ [\ \Lambda \mid N_1, N_2, \cdots, N_{10}] \approx [0.0998,\ 0.0776,\ 0.8226]. \tag{23}$$

and so the final premium is 271.2 Euros. Replacing the final 0 in the history with 10 ($\mathcal{H} = [0, 0, 1, 0, 0, 0, 0, 0, 1, 10]$). The premium for a driver with a large amount of claims can be calculated, the posterior is then

$$\mathbb{P}\ [\ \Lambda \mid N_1, N_2, \cdots, N_{10}] \approx [0.9829,\ 0.0160,\ 0.0011], \tag{24}$$

indicating that the risk profile of this driver is very probable to be the same as the trial-licence risk class. The premium is 609.8 Euros. When repeating the calculation for a driver with zero claims in the last ten years ($\mathcal{H} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$), the posterior becomes

$$\mathbb{P}\ [\ \Lambda \mid N_1, N_2, \cdots, N_{10}] \approx [0.0194,\ 0.0327,\ 0.9479], \tag{25}$$

indicating a risk closer to an experienced driver. The premium is then 233.4 Euros. The premiums in the three cases are summarized in table 3.

Claims last 10 years	Premium by European method
0	233.4
2	271.2
12	609.8

Table 3: Summary of the premiums for the three cases using the European method.

4.3.2 American credibility method

Let's briefly recall that according to this method the premium is computed as

$$\text{Premium} = \mathcal{Z}\bar{S}_k + (1 - \mathcal{Z})M \tag{26}$$

where

$$\mathcal{Z} = \frac{-\sqrt{k}rM}{\sqrt{Var(S)}\ \Phi^{-1}\left(\frac{1-p}{2}\right)}. \tag{27}$$

Here $\bar{S}_k$ is the mean of the total claim amount per year for the policyholder considered. The values of the parameters r and p are fixed at 0.05 and 0.9.

We estimated $Var(S)$ from the data of the whole population and in this way the formula for $\mathcal{Z}$ turns out to depend only from k and not directly from the personal history of the policyholder $\bar{S}_k$, that is considered only for the calculation of the final personalized premium (26). The value of $\mathcal{Z}$ is very low also for high values of k (see table 4), that sounds reasonable accordingly to the previous consideration and to the formula for the final calculation.

It's worth remembering that the formula for $\mathcal{Z}$ is an approximation valid only for histories that are large enough so that the results for small k are not so relevant, even if we computed them for completeness. This can be pointed as a relevant limitation.

One could choose to calculate variance in different ways in order to personalize further the credibility factor. A possible alternative choice are the estimation based on the data filtered only for the risk class of the policyholder.

Number of years (k)	**American credibility factor ($\mathcal{Z}$)**
1	0.007363
2	0.010412
3	0.012752
4	0.047250
5	0.016463
10	0.023283
20	0.032927
30	0.040327
40	0.046565
50	0.052061
60	0.057031

Table 4: American credibility values for different numbers of years if $Var(S)$ is estimated from data of the whole population.

In order to compare the American credibility method to the European, we assumed that each claim size in the personal claim history was the average claim size of that risk class. We then computed premiums for three different 10 year long claim histories, which were the same as in the European case (see above). This was done assuming all the three different risk classes. It is though worth noticing that it is not realistic to have a long history for trial license's holders and beginners. For the result, see table 5.

Claims last 10 years	Trial	Beginner	Experienced
0	599.7	350.6	215.9
2	627.7	374.7	240.4
12	767.7	495.1	363.2

Table 5: Premium by past history and initial risk class assessment using the American credibility method.

5 Discussion and conclusions

Three increasingly complex models were suggested to model insurance policy premiums. Firstly, a model for the whole population suggested a fair premium when not taking into account personal data. In the second modelling step, an important finding is that a classification of policy holders into three different risk classes based on driving experience seems both to be significant and sufficient in order to create more specialized risk premiums. As we can see in Table 2 this can be very significant to the insurance company, as it can lower premiums for approximately 9 out of 10 policy holders. This allows them to keep a more competitive price.

In the final modelling step, two methods to take personal history into account were implemented. As illustrated in tables 3 and 5, the methods give quite different results. It is difficult to say which method is the best, but certain advantages and disadvantages are clear. Firstly, using the European method an initial classification of the driving experience is not required, whereas it is for the American method. This is an advantage with the European method, since such a classification might not be available in practise. On the other hand, the European method has a premium that is bounded, which makes it susceptible for very risky individuals. The premium by the American method never reaches a bound. Two advantages with the European method, however, is that is can be considered more explainable and that its assumptions do not fail when a short history is available, like the American method does. In our model clients were only divided based on their driving experience and it makes this third step inapplicable in practice at least for trial license's holders (histories can be available only for people with a minimum of experience). However, the unbalance of data for driving experience leads us to reasonably think that experienced drivers are by far the most common and allows this premium estimation to be applicable to the vast majority of policyholders.

We have observed other possible approaches to form risk classes in the literature (3; 4). One other approach would have been to use some clustering algorithm on all the policy holder features, to obtain risk classes by more complex patterns. The reason to why we chose to not use this approach is firstly that no consequential results were observed when this was tried initially. Secondly, we deemed it important to know that each of the used variables actually have significance in predicting which premiums to use, and furthermore that the risk classes we obtained were actually relevant. As such, a more rigorous approach fitted our needs better.

In other studies regression models have been used either for the whole population or within risk classes to predict premiums. This can yield even more accurate premiums and would perhaps be a good further step in improving the model.

6 Group work dynamics

The group work went very well overall, and we are satisfied with the project as well as our performance. We tried to divide the work as much as possible, and work in parallel when we could. At the same we also discussed things in the whole group when we felt stuck or when we needed to make important decisions.

One challenge we faced was in the second task, when we ran many statistical tests. At some points it was difficult since there were many variables being tested by many tests, and we all had different results on which variables were significant and which were not. We solved it by calmly going through our finding and discussing. Writing on the blackboard was also very beneficial.

In regard to planning the work, we had a clear structure set from the start, since the project was very naturally split into three sequential parts. As such we only had to complete the parts sequentially and discuss with the instructor when we were satisfied with the results.

7 Instructor's assessment

It is a great pleasure for me to provide positive feedback on my student's performance during their recent teamwork at the ECMI modeling week in Szeged during July 2023. Their exceptional work ethic, dedication, and collaborative spirit stood out throughout the project.

None of the students had previous experience with insurance data modeling, or had taken a course in actuarial mathematics, so the presented problem was completely new and unknown to all of them. They have mastered understanding the problem and start with the modeling very fast, always keeping an eye on the interpretation of results and fine-tuning of the model, taking into account not only mathematical/statistical rigor but also fairness as an argument that a real-life actuary would need to find a well-balanced premium for the policyholders.

Their ability to communicate effectively, inspire their team members, and ensure everyone's contribution was equally valued truly impressed me. Their pre-existing knowledge as coming from different universities and different bachelor/master programs turned out to be marvelously complementary. Some of the students were more skilled in Python as a programming language, some in R as a statistical tool, and some more in theoretical probability and theoretical statistics. By the end of the first day they were teaching each other the necessary skills, sharing information, exchanging ideas and discussing all options, questioning the ideas and weighing all the options, and discussing in a similar manner one would hear from a professional researcher group.

Their creative problem-solving skills and attention to detail resulted in a polished final model that fulfilled all my expectations. I am incredibly proud of their outstanding

performance and their positive impact on the entire experience of being an instructor at the ECMI modeling week.

References

[1] Klugman et al., (2012) Loss models: From data to decisions.

[2] Walters, M. A. (1981). Risk Classification Standards. Proceedings of the Casualty Actuarial Society, LXVIII(129), 1-18.

[3] Xie, Shengkun., (2019) Defining Geographical Rating Territories in Auto Insurance Regulation by Spatially Constrained Clustering

[4] Yeo et al., (2001) Clustering Technique for Risk Classification and Prediction of Claim Costs in the Automobile Insurance Industry.

[5] van der Maaten, Laurens., (2008) Visualizing data using T-sne.

[6] Hinton, Geoffrey. Roweis, Sam., (2002) Stochastic Neighbor Embedding

Appendix

Variable	$\rho_{\text{avg claim amount}}$
Policyholder	0.016
Income (EUR)	-0.013
No. of family members	0.018
Mileage (km)	-0.013
Number of claims	0.018
1. claim amount	0.999
2. claim amount	0.102
3. claim amount	0.015
4. claim amount	-0.009
Average claim amount (EUR)	1.0
Sex_female	0.011
Sex_male	-0.011
Driving Experience_trial licence	0.05
Driving Experience_beginner	-0.009
Driving Experience_experienced	-0.032
Driving Experience_senior	0.009
Driving Experience_business	0.005
Education_high school	-0.001
Education_bachelor	0.015
Education_master degree	-0.024
Education_doctoral degree	0.016
Vehicle production year_before 2007	-0.001
Vehicle production year_2008-2012	-0.009
Vehicle production year_2013-2017	0.019
Vehicle production year_2018-2022	-0.009
Vehicle color_black	-0.007
Vehicle color_blue	-0.007
Vehicle color_green	0.023
Vehicle color_grey	-0.004
Vehicle color_purple	0.004
Vehicle color_red	0.003
Vehicle color_white	0.013
Vehicle color_yellow	-0.024
Manufacturer_Audi	-0.011
Manufacturer_BMW	0.004
Manufacturer_Dacia	-0.002
Manufacturer_Fiat	0.019
Manufacturer_Hyundai	0.015
Manufacturer_Opel	-0.017
Manufacturer_Peugeot	-0.018
Manufacturer_Skoda	0.01
Manufacturer_Toyota	-0.01
Manufacturer_Volkswagen	0.009
Territory_A	0.003
Territory_B	-0.008
Territory_C	0.005

Table 6: Correlation values between explanatory variables and the average claim amounts per user after One-hot encoding of the categorical variables

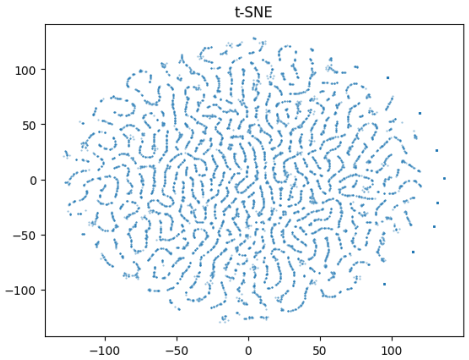

Figure 6: Output of t-SNE for color and the number of claims

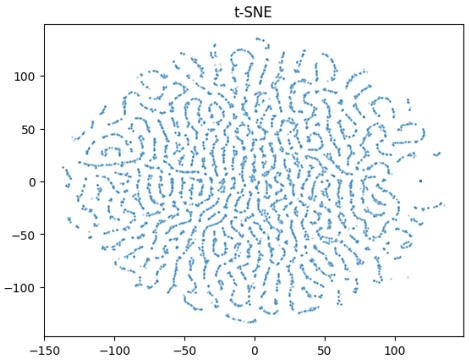

Figure 7: Output of t-SNE for education and the number of claims

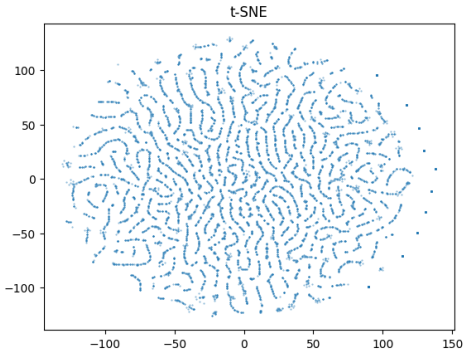

Figure 8: Output of t-SNE for manufacturer and the number of claims

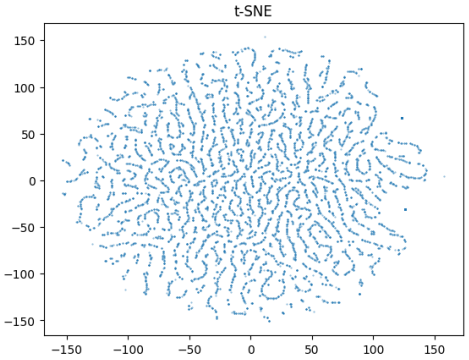

Figure 9: Output of t-SNE for sex and the number of claims

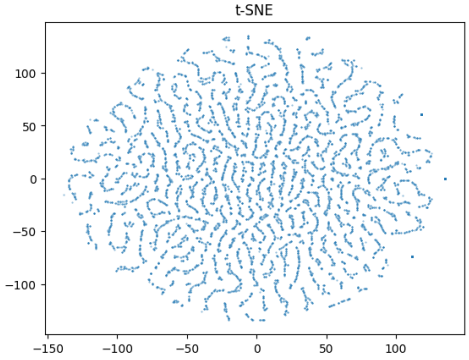

Figure 10: Output of t-SNE for territory and the number of claims