International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00392-x

REGULAR PAPER

A data science approach to risk assessment for automobile insurance policies

Patrick Hosein[1]

Received: 13 September 2022 / Accepted: 5 March 2023

Abstract
In order to determine a suitable automobile insurance policy premium, one needs to take into account three factors: the risk associated with the drivers and cars on the policy, the operational costs associated with management of the policy and the desired profit margin. The premium should then be some function of these three values. We focus on risk assessment using a data science approach. Instead of using the traditional frequency and severity metrics, we instead predict the total claims that will be made by a new customer using historical data of current and past policies. Given multiple features of the policy (age and gender of drivers, value of car, previous accidents, etc.), one can potentially try to provide personalized insurance policies based specifically on these features as follows. We can compute the average claims made per year of all past and current policies with identical features and then take an average over these claim rates. Unfortunately there may not be sufficient samples to obtain a robust average. We can instead try to include policies that are "similar" to obtain sufficient samples for a robust average. We therefore face a trade-off between personalization (only using closely similar policies) and robustness (extending the domain far enough to capture sufficient samples). This is known as the bias–variance trade-off. We model this problem and determine the optimal trade-off between the two (i.e., the balance that provides the highest prediction accuracy) and apply it to the claim rate prediction problem. We demonstrate our approach using real data.

Keywords Risk · Motor insurance · Machine learning · Premium pricing · Claims prediction

Mathematics subject classification 62-08 · 62-11 · 68T09 · 68T99

1 Introduction

Traditionally insurance companies have determined automobile policy premiums using rate tables computed by Actuaries [9]. Today, however, the vast amount of data collected in electronic form can now be used to determine more suitable premiums for a given policy since such data can be used to better predict risk [5]. Furthermore, by using data from present and past customers, the predictions are better suited for the particular environment in which the insurance company operates. This form of personalized policies benefit the customer (who pays an amount more in line with their risk) as well as the insurance company (which can now better ensure that it can safely cover claims costs from risky policies). The typical approach is straightforward. For a given new customer, one can use historical data of past and present customers with similar characteristics (features) to better estimate the risk level of the new customer and then use this to determine a premium for their policy. This is similar to recommender systems used by companies such as Netflix. In that case, movies are recommended to an individual based on movies that were enjoyed by customers with similar characteristics (collaborative filtering). In the case of insurance, one must recommend a policy that is both desirable to the customer (through personalization) and profitable to the insurance provider.

2 Related work and contributions

Many past papers have focused on recommender systems for insurance companies where one of a small number of insurance products is offered. In [17,18], they used historical

✉ Patrick Hosein
patrick.hosein@sta.uwi.edu

[1] Department of Computer Science, The University of the West Indies, St. Augustine, Trinidad and Tobago

Published online: 22 March 2023

data of existing and past customers to determine the most suitable policy for a new customer. In this case, a relatively small number of insurance products are available, and hence, the number of customers who have been using a specific product will be sufficiently high so that the sample size is not an issue when computing the recommendation. The paper [14] also addresses the same problem but focuses on speeding up the computation of the recommendations. The papers [2, 10] do address personalized auto insurance premiums but they focus on using Telematics data to do so. Such devices are not available from all insurance companies and hence have limited applicability. The authors in [6] use fuzzy logic to come up with a rule-based approach to risk. In our case, we use a data science approach and focus on personalization while using traditionally available data.

Several papers also focus on risk assessment. In general, few customers make claims during a year. Furthermore, the claims that are made vary widely from minor incidents (such as a scratched bumper) to major ones (such as when a car has to be written off because it cannot be repaired). This results in a large variation in the average annual claims made by a customer making it difficult to predict. Therefore, papers generally focus on predicting either severity (the expected claim value given that a claim is made, [17,19]) or frequency (the expected number of claims made per year, [3,15]). Another typical metric is the loss ratio which is the ratio of the claims made to the premium charged ([7,20]). We focus on a more direct measure which is the average total value of claims made per year for a given policy (shortened to simply claims rate) which can be thought of as the product of severity and frequency and hence captures both metrics. As mentioned before, predicting claims rate can be challenging because of its high variance. A significant number of samples are required for a good estimate, but as one tries to achieve greater personalization the number of available samples decreases. We investigate the optimal trade-off between these objectives.

Note one potential issue of recommender systems is the following. The recommender system chooses the most appropriate product for the customer, but this may not be a very profitable product for the company and so this trade-off must be taken into account (see [12] for a more detailed discussion on this issue). In our case, we need not worry about this issue since we are focusing on providing the most suitable (unique) product and the premium is then determined to achieve an acceptable profit for each unique policy.

The article [1] summarizes the presentations in the 2018 Swiss Risk and Insurance Forum. In particular, they discussed how data science is being used by actuaries to predict automobile insurance risk. They state that “The current challenge consists in designing and applying effective regression techniques to find an appropriate function that links response variables like the number of claims, the yearly aggregated claim cost, or the remaining time up to the next policyholder event (e.g., claim, lapse) with a highly increasing number of features that include traditional ones and new ones like telematics data, data coming from the Internet of Things, as well as web-based data, including the purchasing strategy of the customer during the acquisition process.” We address this challenge by predicting risk of a new policyholder given demographic and automobile features. Such is the case in developing countries where sensor-based information (telematics and those from Internet of Things) is not easily available. Furthermore, using customer-specific information (such as Telematics) can be considered as a content-based filtering approach recommender system which suffers from the cold-start issue (how do you make a risk prediction for a new customer). Our approach follows a collaborative filtering recommender system approach by using data from other policyholders to make a prediction for a new customer.

Actuaries have begun using more machine learning approaches to risk prediction such as those presented in [4,8]. However, here they focused on claim frequency and severity predictions separately under the assumption of independence. However, [13] demonstrates that this is not necessarily true. We do not make this independence assumption and instead predict total annual claims which takes into account frequency and severity. One issue in jointly considering frequency and severity is the reduction in the samples available for a given Tarif class. Our approach addresses this problem by trading personalization for robustness in such cases.

Instead of a finite number of products (tarif class approach) from which to choose for customer offerings, we provide a unique (personalized) product to each customer. Furthermore, we only take into account demographic and other data collected from each policy but do not consider Telematics data. Naturally our approach can include such data as well and we plan to address this integration in future work. Therefore, our contributions can be summarized as follows:

Claim Rate Prediction: We are predicting annual claim amounts rather than claim frequency or claim severity (which tend to be correlated).

Personalization versus Robustness: We are making a personalized prediction instead of predicting for a Tarif class (in which customers are clustered and all customers within a class are assumed to pose the same risk). Our approach adjusts the degree of personalization based on the number of available historical samples with similar features to those of the test sample.

Feature Importance and Selection: We introduce a method for feature importance determination and this is then used for feature selection.

Interpretability: We demonstrate how the proposed model can be used to explain the basis of the prediction.

This information can then be used to inform the customer of the reasons for the offered premium price.

3 Problem formulation and assumptions

We formulate a model for this problem and then develop an algorithm for its solution. Our objective is as follows: Given policy information for a new or renewal automobile policy (i.e., information about the drivers, cars, etc.), predict the expected total amount in claims that will have to be paid out to this customer over the subsequent year of the policy. This prediction will be based on several factors, but correlates with the risk associated with the drivers and cars on the policy. This information can then be used to determine an appropriate premium for the policy. Traditionally this computation is performed using risk tables but independent of the specific historical data of the company's customers. Here we use historical data of the provider's customers to make the prediction. This is more appropriate since the parameters used in the risk tables may have been developed based on a different customer base (country) and so unsuitable for the one under consideration.

3.1 New versus renewal policies

Note that we need to distinguish between a new policy, for which only customer provided data is available, and renewals, for which information about the customer since the start of their policy is available. We develop a model that can be applied to both new and renewal policies. In the case of renewal policies, the historical data of the policy are included in the training set. The proposed approach therefore automatically includes the past claim information of the policy (since it is now included in the training set). Therefore, we assume that all policies that are at least 1 year old are included in the training set used for parameter determination. In this way, recent information is included in the predictions. Note that this means there is no need for an accident penalty or a no-claim discount since these adjustments are implicit.

3.2 Quantity versus currency of data

The more the data that are used for predictions, the more accurate the prediction. However, as we increase the dataset by going further back in time we will be using outdated information (e.g., automobile models, cost of repairs, etc.). We manage this as follows. As the cost of claims increases (with time), the claim rate of a policy will also increase. The prediction we get from using outdated information will therefore be lower than what would actually occur. We therefore scale predictions as follows. We predict the total claims for the previous year, and we then use a scaling factor to ensure that the total predicted claims equals the total actual claims. This scaling factor is then included when making new predictions. This scale factor computation is repeated every year so that the total predicted claims for the upcoming year will be close to the actual total claims for the year.

3.3 Comprehensive versus Third-Party policies

There are two types of policies, Comprehensive, in which the company has to pay for repairs to the customer's car even if they were at fault, and Third-Party, in which the company only pays for repairs to the other involved party in the accident (i.e., the third party). Note that the risk behavior (and claims requests) of Third-Party versus Comprehensive policy customers may be different, but the approach we use has the ability to extract the relevant information. We therefore make predictions using the combined dataset (i.e., policies of both types), but include the type of policy as a feature. However, we sometimes separate the two cases (i.e., assume an infinite distance between Comprehensive and Third-Party samples) to better illustrate a point. Note that the features for both types of policies are the same except that, for Comprehensive policies, there is also the Sum Insured (based on the value of car) feature. This value is set to zero for Third-Party policies, but the same model can be used for both policy types.

3.4 Multi-car versus single-car policies

For each policy, we must predict the total annual claims for the policy which may have multiple drivers and/or cars. Note that a premium is charged per car and the sum of these forms the policy premium. Our model uses the primary car and primary driver of that car as the sample for that policy (and ignores all other drivers/cars). This means that the prediction is made for a single driver/car pair and this can be repeated for each car on the policy to determine the total claim rate for the policy.

4 Dataset description and preparation

The policy data used for this study span a period of 5 years (2016–2020) and were provided by an insurance company in Trinidad. There were a total of 67,124 policy samples available over this period. Some samples were invalid (e.g., negative claim values) or were incomplete, and after these were removed, we were left with 56,152 samples (i.e, 83.6% of the provided samples). No confidential information is disclosed, and all monetary values are normalized. It consists of data collected from past and existing customers. Each policy record consists of policy information, information for each driver on the policy, information for each vehicle on the policy and information on each claim made on the policy since its inception. Some of this information is not relevant for our purposes (e.g., Vehicle Identification Number) and is ignored. Certain features must be derived from the information provided. For example, the policy lifetime is computed as the difference between the termination date and start date (if terminated) or the difference between the present date and the start date (if currently active). Note that the metric of concern is the average claim rate for a driver/car pair. For each policy, we determine the total value of all claims made (by the primary driver) and divide by the total lifetime of the policy (in years) to obtain the claim rate.

Our objective is to predict the claim rate and use this claim rate to determine a suitable price. In order to do this, we focus only on the primary driver and their associated car for each policy. This happens to be the majority of cases so we do not lose too much information. For this driver, we compute the claim rate based on accidents in which they were involved. We remove features that were mostly empty or corrupt and also placed filters to remove anomalous data. The features that were finally used for the problem are provided in Table 1. POL is the policy number which is used as a unique identifier for the policy. CLR is the average claims per year computed for the primary driver and their associated car for the policy. TOC is the type of policy (customer) which we also use as a feature. SIV is the sum insured value of the primary vehicle of the policy and this value is zero for Third-Party policies. All other features are described in the table.

5 Proposed model

The model we propose is unique in that (a) the metric of concern is claims rate and (b) we use a novel solution approach rather than the traditional approaches. We do not present a full comparison with other machine learning approaches in this paper since our intent is to introduce the model. Future papers will include detailed comparisons with state-of-the-art machine learning algorithms.

5.1 Definition of distance metric

In this section, we describe the approach used for predicting the annual financial claims per year (henceforth called claim rate) for a given policy. We denote the set of features that we consider by the set $\mathbf{F}$. Features include information such as age and gender, as well as information about their associated vehicle such as model and body type. We denote the set of samples by $\mathbf{S}$ where a sample is a policy and includes features for the associated driver/car pair. One way to predict the claim rate is to find the expected value of the claim rates of all existing policies with identical features. However, there may be none or very few of such policies. We must therefore include policies with features that are nearby and include them in the average.

In order to find "close" policies, we need to define a distance metric between pairs of categories of a given feature and then use some measure (e.g., Euclidean distance) to define the distance between two policies. We define this distance as follows. For each category v of feature f, let $C(f, v)$ denote the claim rate averaged over all policies that has a value v for feature f. For example, for the feature gender ($f = gender$) with members m and f, let $C(gender, male)$ denote the average claim rate over all male drivers and let $C(gender, female)$ denote the average claim rate over all female drivers. We define the distance between these two categories of this feature by $|C(gender, male) - C(gender, female)|$. In general, if we had several feature categories then the distance between any two of them will be computed in this manner. Therefore, if the test policy has a male driver then their gender distance from another policy with a male driver is 0 while for a female driver it would be $|C(gender, male) - C(gender, female)|$. Note that the same computation is done for numerical features such as age. For example, the distance between a 48-year-old and a 30-year-old is given by $|C(age, 48) - C(age, 30)|$. By doing this, we maintain the same measurement unit (claim rate) for all distances. If the 48-year-old is a male and the 30-year-old is a female, then the Euclidean distance is used (i.e., the root of the sum of the squares of the gender and age feature distances).

5.2 Claim rate prediction

If there were several existing policies with the exact feature values as the test policy, then one could obtain a good estimate on the claim rate for the test policy by taking the average of claim rates over all policies with the same features. However, in general there may not be sufficient samples (or none) to obtain an estimate with sufficient confidence, and so, we need to include nearby samples as well. The more the nearby samples we use, the more robust the estimate but the less personalized since included samples are further away. This in turn leads to lower prediction accuracy. We take a weighted average of claims of all policies where the weight is inversely proportional to the Euclidean distance between the policies. Note that other distance metrics can be used. For example, we tried the Manhattan distance metric (instead of Euclidean) since we thought computations would be faster. However, both performance results and computation run times were similar, so we decided to use the more common Euclidean distance metric.

Suppose we wish to predict the claim rate for some test policy and denote the distance between this policy and some training policy s by d_s. We use a weight $(1 + d_s)^{-\kappa}$ for $\kappa \geq 0$ when taking into account the claim rate of sample

Table 1 Policy features used for analysis

Feature	Description
POL	Policy identification number
CLR	Annual claim rate (total claims divided by policy lifetime)
ADR	City of home address
COV	Were drivers continuously insured over the last 5 years? (y/n)
SEX	Gender of driver
AGE	Age of driver
MST	Marital status of driver
USE	Type of use (business, work or pleasure)
WRK	Whether car is used for work (y/n)
NAF	The number of at-fault accidents over the last 5 years
DAF	Number of years primary driver has been free of Claims
NNF	The number of not-at-fault accidents over last 5 years of driver
MAK	Car manufacturer
VYR	Model year
BDY	Body type
YCF	The number of years the primary car has been claim free
NCC	Engine size of car (in CC)
TOC	Type of policy (Comprehensive or Third-Party)
SIV	Sum insured value

$s \in \mathbf{S}$. However, we need to have a normalizing factor α. The predicted claim rate c for the test sample is therefore given by

$$c(\kappa) \equiv \sum_{s \in \mathbf{S}} \alpha \frac{c_s}{(1+d_s)^\kappa} \tag{1}$$

where c_s is the claim rate of policy s. If all policies had the same claim rate, then the predicted claim rate should also have this value, and hence, we must have

$$c \equiv \sum_{s \in \mathbf{S}} \alpha \frac{c}{(1+d_s)^\kappa} \tag{2}$$

and hence

$$\alpha = \left(\sum_{s \in \mathbf{S}} \frac{1}{(1+d_s)^\kappa} \right)^{-1} \tag{3}$$

and so we have the predicted claim rate for the test policy as

$$c(\kappa) \equiv \frac{\sum_{s \in \mathbf{S}} \frac{c_s}{(1+d_s)^\kappa}}{\sum_{s \in \mathbf{S}} \frac{1}{(1+d_s)^\kappa}} \tag{4}$$

The pseudo-code for this computation is provided in Algorithm 1.

5.3 Computing the optimal value of κ

Next we determine the optimal value of κ. For an existing policy s, we have the actual claim rate c_s. Note that we can predict a claim rate for this sample (in which case the sample must be removed from the training set) and we denote this predicted value by $\hat{c}(\kappa)$. We introduce the hat to distinguish this predicted value with the actual value (which has no hat). Note that we use fivefold cross-validation, and hence, 80% of the samples are used for training (computing the average claim rates $C(f, v)$) while the other 20% are used for testing (and determination of the accuracy). We also tried using tenfold cross-validation and obtained similar performance results but at the cost of greater computational run time so used fivefold for the entire study. Note that when $\kappa = 0$ then $\hat{c}(\kappa) = \bar{c}$, and so, the prediction is simply the average over all (training) samples. As κ is increased, close samples are weighted more heavily but the average becomes less robust, and hence, the error will eventually start increasing again. Therefore, the optimal κ lies somewhere in between (see Fig. 4 for an example of this relationship). We therefore will find κ that minimizes the mean absolute error (MAE) of the

Algorithm 1 Pseudo-code for proposed Algorithm to predict test sample claim rate $c(\kappa)$

```
$\mathbf{F} \equiv$ set of features
$\mathbf{S} \equiv$ set of training samples
$\mathbf{v}_f \equiv$ set of categories for feature $f \in \mathbf{F}$
$\kappa > 0$ tuning parameter
$X_{sf} \in \mathbf{v}_f \equiv$ category of feature $f \in \mathbf{F}$ of training sample $s \in \mathbf{S}$
$x_f \in \mathbf{v}_f \equiv$ category of feature $f \in \mathbf{F}$ for test sample
$c_s \equiv$ claim rate for sample $s \in \mathbf{S}$
$c(\kappa) \equiv$ predicted claim rate for test sample using parameter $\kappa$
$\bar{c} \leftarrow \frac{1}{|\mathbf{S}|} \sum_{s \in \mathbf{S}} c_s$ (average claim rate over all training samples)
for each $f \in \mathbf{F}$ do
    for each $v \in \mathbf{v}_f$ do
        $\mathbf{z} \equiv \{s \in \mathbf{S} \mid X_{sf} = v\}$
        $C(f, v) \leftarrow \frac{1}{|\mathbf{z}|} \sum_{s \in \mathbf{z}} c_s$ (average over samples where $f$ has value $v$)
    end for
end for
for each $s \in \mathbf{S}$ do
    $d_s \leftarrow \left( \sum_{f \in \mathbf{F}} \left( C(f, x_f) - C(f, X_{sf}) \right)^2 \right)^{\frac{1}{2}}$
    $d_s \leftarrow \frac{d_s}{\bar{c}}$
end for
$c(\kappa) = \frac{\sum_{s \in \mathbf{S}} \frac{c_s}{(1+d_s)^\kappa}}{\sum_{s \in \mathbf{S}} \frac{1}{(1+d_s)^\kappa}}$ (predicted claim rate for test policy)
```

prediction. For convenience, we will normalize this by the MAE if one used the average claim rate over all policies, $\bar{c}$, as the predictor. One can think of this case as making a prediction without features. Therefore, we will compare the error of the prediction made with features with the error of the prediction made without features. Let us denote the test set by $\mathbf{T}$, then we compute the normalized error over the test samples as

$$E(\kappa) = \frac{\sum_{t \in \mathbf{T}} |\hat{c}_t(\kappa) - c_t|}{\sum_{t \in \mathbf{T}} |\bar{c} - c_t|} \tag{5}$$

If the predictor is the same as averaging over all policies (i.e., $\hat{c}_t(\kappa) = \bar{c}$), then this ratio is 1. However, if, by adding features, the MAE of the predictor is decreased then this ratio drops below 1. Therefore, this metric provides an indication of prediction performance using features when compared to prediction performance without using features and hence demonstrates the benefit of the feature-based approach. We then find the κ value that optimizes the predictor as

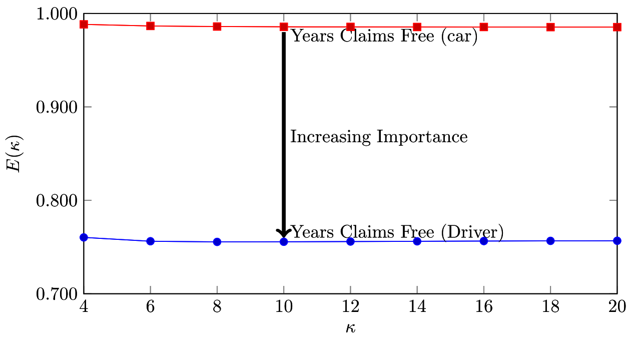


Fig. 1 $E(\kappa)$ of two features to demonstrate relative importance

$$\kappa^* = \underset{\kappa}{\text{argmin}}\, E(\kappa) \tag{6}$$

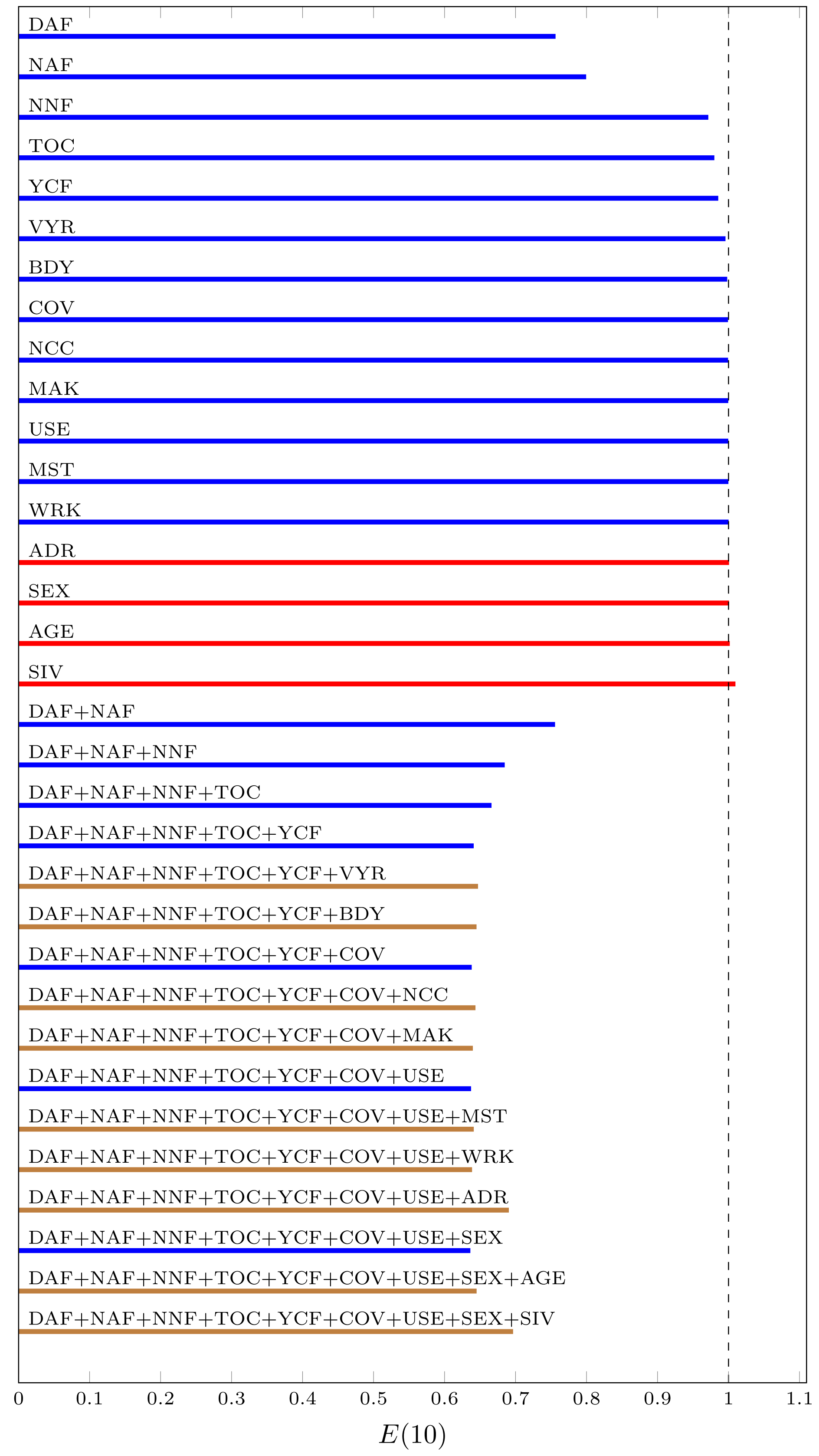
This value is then used to obtain the optimal prediction as $c_t^* = \hat{c}_t(\kappa^*)$.

5.4 Feature importance

Consider a single feature. We know from the previous section that as κ is increased then $E(\kappa)$ should initially decrease before increasing once again. If this does not occur, then the feature does not capture sufficient information to be useful for predictions. One can therefore use the value of $E(\kappa)$ evaluated at the optimal κ for that feature alone as an indication of importance. In fact, even if we used a fixed value of κ for each feature the corresponding value of $E(\kappa)$ is an indication of relative importance with lower values indicating more importance. For example, in Fig. 1 we plot $E(\kappa)$ as a function of κ for two features DAF (years claim free for the driver) and YCF (years claim free for the car). For DAF, the minimum error occurs at $\kappa = 8$ while for YCF it occurs at $\kappa > 20$. However, at $\kappa = 10$ we find that the respective values provide a good representation of the optimal value and hence can be used to compare the two features. Also note that here we clearly see that risk depends primarily on the driver with the car playing a minor role.

We therefore use this approach to determine which features are important and hence should be included in the analysis. We use a value of $\kappa = 10$, and using a single feature at a time, we compute $E(10)$. The resulting values are provided in Fig. 2. The features represented in red have normalized errors greater than 1.

Fig. 2 Normalized MAE for proposed predictor computed for each feature



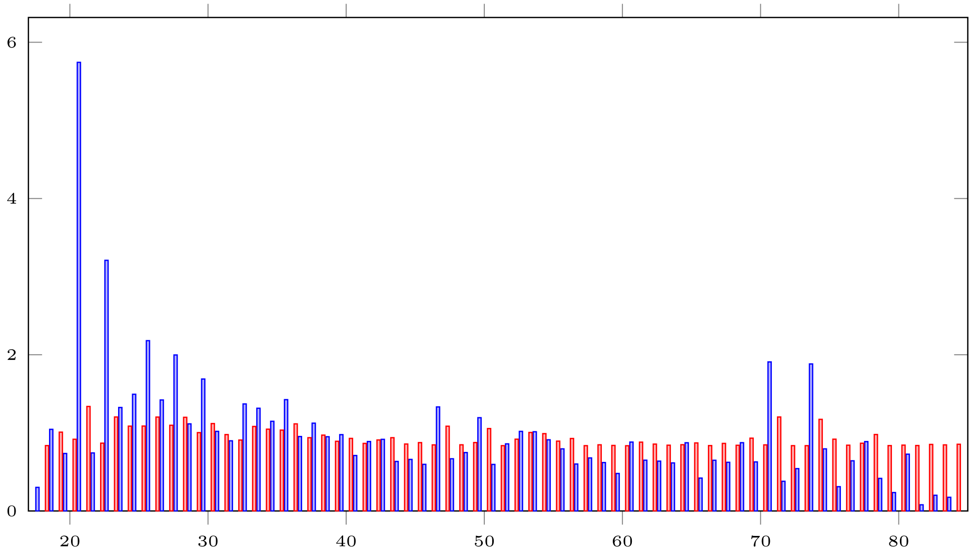


Fig. 3 Histogram of claim rate versus age for original (blue) and filtered (red) cases

5.5 Feature selection

Now that we know which features are important, we focus on which of them should be included in the model. We do this as follows. Starting from the most important feature (lowest value for $E(10)$), we add one feature at a time and again compute $E(10)$ for the combination. If the performance metric decreases (i.e., better results), then we keep it and repeat. If the performance metric increases, then we remove the recently added feature and repeat. Note that some features may have low importance when considered in isolation, but together with other features (such as type of policy) their value increases. The results of this process are provided in the lower part of Fig. 2. A brown bar indicates that the addition of a feature resulted in a loss of performance, and hence, the feature should be removed.

The final features to be used include number of years since the driver was last in an accident, the number of at-fault accidents by the driver over the last 5 years and the number of not-at-fault accidents over the last 5 years. Each of these is a strong indicator of risk. In the case of renewals, we would have actual claim rate values, but for new customers these three features (even without financial information) correlate well with claim rate. The type of policy feature is also needed since it helps to distinguish the two types of policies. Note that we could do the analysis separately for each type of policy, but the increase in sample size by combining the two types provides better overall results. Only one car feature was found to be sufficiently beneficial and that was the number of years since the last claim was made on the car. However, this feature is far less important than the driver features that were included, indicating that what really matters is the driver on the policy and not the car. Whether the driver was continually insured over the last 5 years (i.e., mature driver), the type of use (personal versus business) as well as the gender of the driver were also found to be useful (but far less so than the others).

Note that we had expected certain features (like age) to be beneficial, but they were not. In Fig. 3, we provide a histogram of the average claim rate by age (in blue). We see that there is a weak dependency on age, but because of the large variations from year to year (because of limited data), the dependency is not sufficiently robust. Next we predicted the claim rate for each age using the approach described previously. We found that the optimal value of κ was 2 with a normalized MAE of $E(2) = 0.9996$ which indicates that limited personalization was possible. We then used this value to find optimal claim rate values for each age. In Fig. 3, we provide the histogram of the original claim rates (blue) and the filtered claim rates (red). We note that the red claim rates are each close to unity and hence provide little differentiation. This is why this feature does not provide much benefit for predictions.

5.6 Parameter optimization

We now have the set of features to be included in the model. Next we find the optimal value of κ for this combination of features. This value will then be used for making predictions. In Fig. 4 (brown curve), we plot $E(\kappa)$ as a function of κ. We find the optimal value to be $\kappa^* = 8$ with $E(8) = 0.63$, and hence, one can reduce the MAE obtained with no features by 37% by using the 8 chosen features. We also note that, although E increases with κ beyond the optimal point, the increase is very gradual, so the error remains nearly constant for a wide range of values, and so, the approach is robust with respect to κ.

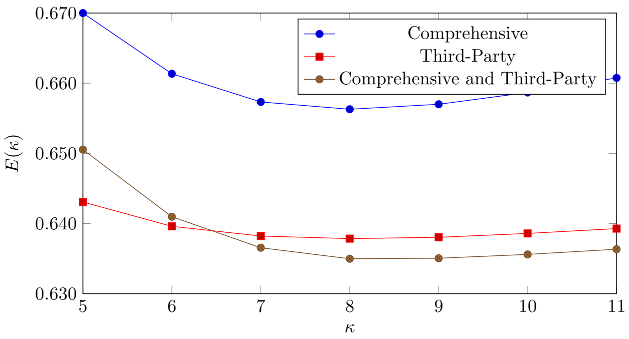


Fig. 4 $E(\kappa)$ as a function of κ for selected features

We believe that if we had performed feature selection using each policy type separately that we would get the same features. We therefore used these features and determined $E(\kappa)$ for Third-Party policies only and also for Comprehensive policies only. These are also plotted in Fig. 4. We find that the accuracy for Third-Party only samples is close to that of the case of using both Third-Party and Comprehensive samples. This is primarily due to the fact that there are 50% more Third-Party samples than Comprehensive samples. Therefore, the Third-Party samples are more useful to the Comprehensive predictions than the other way around. Also note that all three cases are optimal at $\kappa = 8$.

6 Illustrative examples of predictions

We have now determined the features to be used and the optimal value of κ. In this section, we will consider various policy scenarios and predict the resulting claim rate to demonstrate the dependence on the features. Although we use both Comprehensive and Third-Party policies in our model, we will illustrate using a Comprehensive policy and normalize claim rates with respect to the average over all Comprehensive Policies. In Table 2, we provide various scenarios to illustrate that the model provide reasonable outputs. The top table starts with a low-risk policy and features are changed one at a time that results in increased claim rates. The bottom table starts with a high-risk policy and features are adjusted one at a time in order to lower the claim rate.

There is one outstanding case that provided unexpected results. In the lower table, when we reduce the number of not-at-fault accidents from 1 to zero we expect a decrease in the claim rate, but instead we found that it increased. We investigated this in detail. We found that the provided data has some inconsistencies. There were many cases where the number of not-at-fault accidents and at-fault accidents were 0, but the driver indicated that they made a claim within the last 5 years. This of course is inconsistent. This lead to a lot of claims listed under NNF = 0 when they should be listed under NNF = 1 or above. We believe this to be the reason for the result obtained. Our intent was not to make any adjustments to the given data for this paper to avoid any appearance of data tweaking. However, in the future, we will investigate what happens when we adjust the data to make it more consistent while justifying any changes made.

7 Interpreting prediction

The ability to determine the features that played a dominant role in the prediction of risk of a customer is important. Since pricing depends on this risk, then one should be able to explain to a new customer what factors played a role in the determination of their policy premium. This is often difficult for certain machine learning algorithms although there are proposed ways to address this issue [16]. In this section, we provide an approach for interpreting results using the model that was used for predicting risk. In future work, we will compare with the approach specified in [16] which uses Shapely values.

Once a predicted claim rate is computed, then this information can be used to compute a premium. The premium will take into account the operational costs of the company as well as the desired profit margin. This is another interesting area of research, but is outside the scope of this paper. Once a premium is computed, it is important to explain the reason for the amount (i.e, interpretability). The operational cost and profit are independent of the customer so the only customer-dependent factor is the predicted claim rate. We can determine the influence of each feature on this claim rate and this information can be used to explain the decision made. We do this as follows. Consider any feature f and let v represent the category value of this feature for the new policy. We can use the model, with only feature f, to determine the predicted claim rate for anyone in category v. Let us denote this predicted claim rate of this feature by $\tilde{C}(f, v)$. Note that this is not the same as the average claim rate over all training samples with category value v which we previously denoted by $C(f, v)$. Let us explain with the feature gender. If $\kappa = 0$ then $\tilde{C}(\text{gender}, \text{male}) = \tilde{C}(\text{gender}, \text{female}) = \bar{c}$. However, as κ goes to infinity then $\tilde{C}(\text{gender}, \text{male})$ approaches $C(\text{gender}, \text{male})$ and $\tilde{C}(\text{gender}, \text{female})$ approaches $C(\text{gender}, \text{female})$. For positive values of κ, $\tilde{C}(f, v)$ will lie between $\bar{c}$ and $C(f, v)$.

The metric $I_f \equiv \tilde{C}(f, v)/\bar{c}$ will be used to represent the impact of feature f (of a policy with value v for the feature) where $\bar{c}$ is the average claim rate. For this exercise we only use Third-Party samples to better explain the approach. If $I_f < 1$ then the feature is causing a reduction of the claim rate otherwise it is causing an increase in the claim rate. Note that all values are being computed using $\kappa = \kappa^*$ and normalized with respect to the average claim rate for Third-Party policies.

Table 2 Predictions for sample cases starting with low-risk case (top) and high-risk case (bottom)

DAF	NAF	NNF	TOC	YCF	COV	USE	SEX	Claim rate
15	0	0	CM	15	Y	Business	M	0.07
6	0	0	CM	15	Y	Business	M	0.63
2	1	0	CM	15	Y	Business	M	0.93
15	0	1	CM	15	Y	Business	M	0.68
15	0	0	TP	15	Y	Business	M	0.03
15	0	0	CM	0	Y	Business	M	0.20
15	0	0	CM	15	N	Business	M	0.07
15	0	0	CM	15	Y	Private	M	0.06
15	0	0	CM	15	Y	Business	F	0.07
1	1	1	CM	1	N	Private	F	6.04
5	1	1	CM	1	N	Private	F	1.53
6	0	1	CM	1	N	Private	F	0.56
1	1	0	CM	1	N	Private	F	7.59
1	1	1	TP	1	N	Private	F	4.04
1	1	1	CM	15	N	Private	F	5.55
1	1	1	CM	1	Y	Private	F	5.97
1	1	1	CM	1	N	Business	F	6.00
1	1	1	CM	1	N	Private	M	6.04

For the policy we chose, we have the following information. The driver got into an accident and made a claim 9 years ago ($I_{DAF} > 1$). They have no at-fault accidents over the last 5 years ($I_{NAF} < 1$). They have had 1 not-at-fault accidents over the last 5 years ($I_{NNF} > 1$). They have a Third-Party Policy (hence $I_{TOC} = 1$). Their car was last in an accident 18 years ago ($I_{YCF} < 1$). They have been continuously insured over the last 5 years ($I_{COV} < 1$). This is their private vehicle ($I_{USE} < 1$). The driver is Male ($I_{SEX} < 1$ but almost 1). The predicted claim rate (which was impacted by the various features) has $I_{CLR} = 1.0$. We provide this information visually in Fig. 5. Hence, the provider can explain to the customer the specific reasons for the premium of their policy.

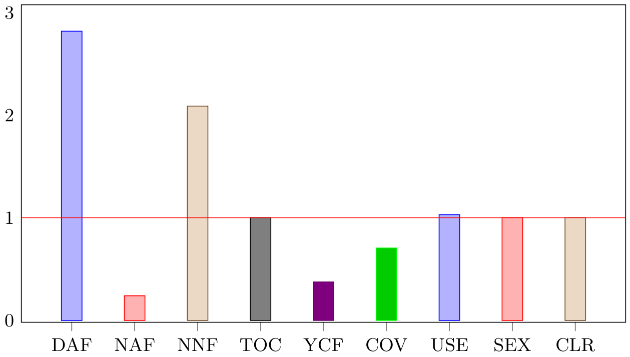


Fig. 5 Contribution of each feature to prediction

8 Some analytical results

8.1 Expected value of prediction

Consider the predicted claim rate $\hat{c}$ for some test sample. If we assume that the training samples have an average claim rate $\bar{c}$, then we will show that the expected value of $\hat{c}$ is $\bar{c}$. This would mean that the sum of predicted claim rates approaches the sum of the actual claim rates as the number of test samples increases. This ensures that, at the end of the year, the total claims that are predicted is close to the total actual claims that were made.

Lemma 1 *If the training samples have a mean claim rate of $\bar{c}$, then the expected value of the prediction for a test sample is equal to $\bar{c}$.*

Proof Recall that the predicted value for a given κ is given by

$$\hat{c}(\kappa) = \frac{\sum_{s \in \mathbf{S}} \frac{c_s}{(1+d_s)^\kappa}}{\sum_{s \in \mathbf{S}} \frac{1}{(1+d_s)^\kappa}}$$

We need to take the expectation of the right-hand side. Now note that the feature values for a particular set of training samples are fixed and only the claim rate varies. Also note that the feature values of the test sample is also fixed. This means that d_s (for sample s) is fixed given the specific test and training samples, and so, the denominator is constant, and hence, we have

$$\mathcal{E}[\hat{c}(\kappa)] = \frac{\sum_{s \in \mathbf{S}} \frac{\mathcal{E}[c_s]}{(1+d_s)^\kappa}}{\sum_{s \in \mathbf{S}} \frac{1}{(1+d_s)^\kappa}}$$

Using the fact that $\mathcal{E}[c_s] = \bar{c}$, we obtain $\mathcal{E}[\hat{c}(\kappa)] = \bar{c}$. □

8.2 Limiting values of $c(\kappa)$

We have seen that $\hat{c}(\kappa = 0) = \bar{c}$. In this section, we compute the limit of $\hat{c}(\kappa)$ as κ tends to ∞. We then take the expected value of this limit.

Lemma 2 *If the training samples have a mean claim rate of $\bar{c}$, then*

$$\mathcal{E}[\lim_{\kappa \to \infty} c(\kappa)] = \bar{c} \tag{7}$$

Proof We first compute the limit of the predicted value, $\hat{c}(\kappa)$ as κ tends to infinity. There are two cases to consider. If one or more training samples have identical feature values as the test sample (i.e., $d_s = 0$), then $\hat{c}_s$ is the average of these values. However, if none of the training samples have identical features then both the numerator and denominator of $\hat{c}(\kappa)$ tend to zero as κ tends to infinity, so we instead do the following. Denote the training sample with the smallest distance from the test sample by s'. Let us multiply top and bottom of the equation for $\hat{c}(\kappa)$ by $(1 + d_{s'})^\kappa$ to obtain

$$\lim_{\kappa \to \infty} c(\kappa) = \lim_{\kappa \to \infty} \frac{c_{s'} + \sum_{s \in \mathbf{S} | s \neq s'} c_s \left(\frac{1+d_{s'}}{1+d_s}\right)^\kappa}{1 + \sum_{s \in \mathbf{S} | s \neq s'} \left(\frac{1+d_{s'}}{1+d_s}\right)^\kappa} \tag{8}$$

Now note that since $d_{s'} < d_s$ for all samples s, then, as κ goes to infinity, the summations go to zero, and hence,

$$\lim_{\kappa \to \infty} \hat{c}(\kappa) = c_{s'} \tag{9}$$

If multiple samples are at this distance $d_{s'}$, then the numerator constant would be the sum of these samples and the denominator would be the number of them, and hence, the limit is the average of the claim rates of these samples. Since $c_{s'}$ is a sample from the training set space, then its expected value is $\bar{c}$, and hence,

$$\mathcal{E}[\lim_{\kappa \to \infty} \hat{c}(\kappa)] = \bar{c} \tag{10}$$

□

One should note the following. As the number of training samples increases, more training samples will be available close to the test sample, and hence, the optimal value of κ will increase. In the limit, the predicted value becomes the true mean for the features of the test sample, thus achieving precise personalization.

9 Conclusions and future work

We presented a new model for automobile insurance risk assessment and demonstrated its effectiveness using real data. We showed how feature importance can be computed, how features can be selected and how model parameters are optimized. Finally we demonstrated how the model can be used in practice and results interpreted. Note that this approach can be applied to any regression problem and its performance will improve as the variance of the target metric decreases.

There are two differences with prior work: the metric considered (claim rate) and the regression approach used. We plan to isolate the contributions of each of these. We plan to use traditional machine learning techniques with the "claim rate" metric and demonstrate the improvement. We also plan to compare our proposed regression approach with the many machine learning algorithms over a wide range of datasets. One of the problems with the proposed approach is the computation resources required to optimize over κ. We are investigating ways to speed this up mathematically (e.g., a smarter search over κ) and computationally (i.e., through parallel computations). We also plan to investigate how recursive improvements of distance values can be used to increase accuracy (see [11] for an example of this improvement). Finally we plan to investigate properties of the regression model such as whether $E(\kappa)$ is convex when its gradient at $\kappa = 0$ is negative.

Author Contributions The sole author performed the research, wrote the code for evaluating the solution and wrote the entire paper

Funding The authors did not receive support from any organization for the submitted work.

Availability of data and materials The data used for this publication are confidential, and hence, we are only permitted to provide results but cannot share the data.

Code Availability The code used to generate results is also proprietary to the company, but we hope that our pseudo-code can be used if one wishes to apply the model to their datasets.

Declaration

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Albrecher, H., Bommier, A., Filipović, D., et al.: Insurance: models, digitalization, and data science. Eur. Actuar. J. **9**, 349–360 (2019)
2. Bian, Y., Yang, C., Zhao, J.L., et al.: Good drivers pay less: a study of usage-based vehicle insurance models. Transp. Res. A: Policy Pract. **107**, 20–34 (2018). https://doi.org/10.1016/j.tra.2017.10.018
3. David, M., Jemna, D.V.: Modeling the frequency of auto insurance claims by means of poisson and negative binomial models. Analele stiintifice ale Universitatii “Al I Cuza” din Iasi Stiinte economice/Scientific Annals of the“ Al I Cuza” (2015)
4. Denuit, M., Trufin, J.: Effective Statistical Learning Methods for Actuaries. Springer Actuarial Lecture Notes (2019)
5. Errais, E.: Pricing insurance premia: a top down approach. Annals of Operations Research, pp. 1–16 (2019)
6. Esfandabadi, Z.S., Ranjbari, M., Scagnelli, S.D.: (0) Prioritizing risk-level factors in comprehensive automobile insurance management: A hybrid multi-criteria decision-making model. Glob. Bus. Rev. https://doi.org/10.1177/0972150920932287,
7. Guelman, L.: Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. **39**(3), 3659–3667 (2012)
8. Hanafy, M., Ming, R.: Machine learning approaches for auto insurance big data. Risks **9**(2), 42 (2021)
9. Hassani, H., Unger, S., Beneki, C.: Big data and actuarial science. Big Data Cogn. Comput. **4**, 40 (2020)
10. He, B., Zhang, D., Liu, S., et al.: Profiling driver behavior for personalized insurance pricing and maximal profit. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 1387–1396. https://doi.org/10.1109/BigData.2018.8622491 (2018)
11. Hosein, P.: On the prediction of automobile insurance claims: the personalization versus confidence trade-off. In: 2021 IEEE International Conference on Technology Management, pp. 1–6. Operations and Decisions (ICTMOD), IEEE (2021)
12. Hosein, P., Rahaman, I., Nichols, K., et al.: Recommendations for long-term profit optimization. In: ImpactRS@ RecSys (2019)
13. Jeong, H., Valdez, E.A.: Predictive compound risk models with dependence. Insurance Math. Econom. **94**, 182–195 (2020)
14. Kanchinadam, T., Qazi, M., Bockhorst, J., et al.: Using discriminative graphical models for insurance recommender systems. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 421–428 (2018). https://doi.org/10.1109/ICMLA.2018.00069
15. Liu, Y., Wang, B.J., Lv, S.G.: Using multi-class adaboost tree for prediction frequency of auto insurance. J. Appl. Finance Bank. **4**(5), 45 (2014)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., et al. (Eds.) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc (2017). https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
17. Qazi, M., Fung, G.M., Meissner, K.J., et al.: An insurance recommendation system using bayesian networks. In: Proceedings of the Eleventh ACM Conference on Recommender Systems. Association for Computing Machinery, New York, NY, USA, RecSys ’17, pp. 274–278 (2017). https://doi.org/10.1145/3109859.3109907
18. Qazi, M., Tollas, K., Kanchinadam, T., et al.: Designing and deploying insurance recommender systems using machine learning. WIREs Data Min. Knowl. Discovery **10**(4), e1363 (2020). https://doi.org/10.1002/widm.1363
19. Su, X., Bai, M.: Stochastic gradient boosting frequency-severity model of insurance claims. PLoS ONE **15**(8), e0238000 (2020)
20. Zhang, Y., Dukic, V.: Predicting multivariate insurance loss payments under the bayesian copula framework. J. Risk Insurance **80**(4), 891–919 (2013)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.