Insurance Claims Risk Modeling and Forecasting Using Mathematical Models and Scorecards

Nataliia V. Kuznietsova[1,2], Illia O. Kvashuk[1] and Anna O. Chemanova[1]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", ave. Beresteiskyi 37, Kyiv, 03056
[2] Claude Bernard Lyon 1 University, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex

Abstract

In this paper, several car insurance claims problems are analyzed and solved via existing statistical models implementation for real-world datasets. The first problem which was studied is the problem of measuring the probability of a claim for a specific policy. This problem is solved by using a set of families of generalized linear models with an additional approach to analyze data by utilizing survival models. The best generalized linear model is then chosen according to statistical criteria. The second problem considers distinct classes of policies. A number of claims and prices are forecasted for the different groups. Same approach as for the first problem, generalized linear models are used and the best model is chosen according to statistical criterion. The third problem is the problem of scorecard generation. A brief interpretation and result of the built scorecard is also provided.

Keywords

Car-insurance 1, Generalized linear models 2, Scorecard 3, Survival models 4, Claims forecasting 5

1. Introduction

Usually, the insurance activity is aimed to protect the property interests of individuals and legal entities in the event at the expense of monetary funds, which are formed from the insurance premiums paid by policyholders. One of the main conditions for the effective functioning of the insurance market is the reliability of its participants – insurance companies. Supporting the ability of insurance companies operating in the market to fulfill their obligations promptly and as a whole. That is their financial stability which is a special starting point for the actual manifestation and implementation of the insurance function. The current financial state of the insurance companies requires the search for new forms and methods of increasing their competitiveness and financial stability. They need to create special decision-support systems for more effective assessment of the policies, more precise forecasting of the probability of claims, evaluate the possible losses and develop more flexible conditions for insurance policy evaluation.

The variety of risk manifestation forms and the frequency and complexity of the consequences of their implementation determine the need for an in-depth analysis of possible risks and economic-mathematical justification of the financial policy of insurance companies. For every car insurance company importance of proper policy selection for a given client cannot be overestimated. The insurance premium is formed according to the client's expectation to be prone to raise claims and the size of those claims [1–3].

Information for the determination of terms and conditions of policies can be separated into two parts: data concerning a driver and a car. Age, driving, and length of insurance policy are the values that define a driver part of the information. However, some aspects like driver's habits are hard to collect, describe and analyze. On the other hand, information about cars can be specified and collected

ITS-2023: Information Technologies and Security, November 30, 2023, Kyiv, Ukraine
natalia-kpi@ukr.net (N. Kuznietsova); illiakvashuk@gmail.com (I. Kvashuk);
ankachemanova@gmail.com (A. Chemanova)
0000-0002-1662-1974 (N. Kuznietsova); 0009-0006-5585-3045 (I. Kvashuk);
0009-0008-6647-9364 (A. Chemanova)

concerning some technical criteria [2]. It can range from the car type to a quantity of cylinders or safety bags. A practical task is to model and forecast claims with information about cars being available in abundance, hence requiring selections and filtering in search of its most relevant parts.

A completely separate issue is creating models that can be used to predict some aspects of a claim based on selected data. One of the most important tasks is to predict the probability of a claim for a specific case. Insurance firms need to have a proper model for predicting and forecasting claims for different clients. Meanwhile, those methods should be easily interpreted and thus explained to clients or regulators about key factors that affect the terms and conditions of policies.

2. Problem statement

This work is concentrated on solving the main problems, which appear in the insurance field. The first and foremost task is that the claim expectation should be forecasted for a given client. It could be measured by the claim's probability. Companies need a way to approximate the chances of claims to properly form policies' terms for a given client. This task requires taking into account the client's data and forming a decision based on it.

The second task is forecasting the number of claims for each group. The importance of this task is quite understandable while it is a part of company policy selection. By grouping clients by aggregating values, groups can be created. For these groups, the number of claims can be estimated and the models for forecasting can be built. The approach can follow two possible scenarios: modeling only the number of claims or total spending on a group.

Third task the model creation, which is usually paired with interpretation. This interpretation can provide valuable insight into what values increase the probability of the claim. This allows us to create scorecards that can be built to provide an easy tool to make decisions directly from data provided by a client. The main objective of this study is to define not only the probability and cost (value) of each claim but also the subset of the most damaged cases.

3. Methods

The appropriate approach usually depends on the task but the most important is that it is determined by the flow of data extraction and preparation. The same method can be applied to the same data but different approaches and pre-processing techniques may affect the results. For example, [3] provides us with the flow and handling of data and objectives very similar for use in this work. Data is collected on an open platform. Claims are analyzed and the number of which match our task is predicted. However, due to the dataset restriction, the preprocessing was added which yielded comparatively stuffiest results but lacked interpretability due to PCA usage.

3.1. Generalized linear models

Generalized linear models (GLM) were the main tools used during our research. They provided a unified framework for modeling and forecasting the target variables [3]. Due to the variable's nature and the different tasks that were tackled, the number of family distributions was used to deal with the problems from different sides and selections of the most fitting.

The generalized linear model is an extension of a simple linear regression model. A linear relationship between variables is the simplest case for researching links between factors. However, this is not true for most real-world processes where the relationship is more complicated than linear. In this case linking function is introduced. There are a number of different families that were used in the research.

The general way of writing down the generalized linear model is as follows:

$$X\beta = g(\mu),$$

where X denotes the independent variables and β is a parameters vector g is a link function to transform the scale of dependent variable μ to suit a linear relationship.

Generalized models can be used for discrete or continuous variables which provides it with a significant advantage.

Logistic regression (LR) is a statistical method that is used for classification values into different categories. In the scope of the research, the logistic regression was used for modeling claim probability for the one police.

$$X\beta = logit(\mu) = \ln\left(\frac{\mu}{1-\mu}\right).$$

Normal or Gaussian generalized model uses an identity link function which is the same as simple linear regression.

$$X\beta = \mu.$$

Poisson regression is a statistical model that is used when the dependent variable is a count of occurrence. Its link function is following:

$$X\beta = \ln(\mu).$$

3.2. Survival models

To predict claims or similar events like death or accidents, survival models can be used. They can be utilized when the outcome can be traced along some period of time [4].

The simplest form of survival model is a table with all events noted with timestamp of occurrences. It may give a significant insight into the time periods when most events occur.

3.3. Scorecards

Scorecards are special tables constructed in a way to provide scores for every feature, summing up the scores for a record, the total points can be estimated. It is possible to move records to one of the preselected categories by assigning levels to the score.

Scorecards are powerful practical tools that can be used to fast identify policies with high risks [4]. Scorecards are built by using Weight of evidence - WoE, Information value - IV, and Population Stability Index – PSI.

$$WoE = \ln\left(\frac{Distribution of Goods}{Distribution of Bads}\right).$$

$$IV = \sum(\%of non-events - \%of events) \cdot WOE.$$

$$PSI = \big(\%\ of\ records\ based\ on\ scoring\ variable\ in\ Scoring\ Sample\ (A) -$$

$$\%\ of\ records\ based\ on\ scoring\ variable\ in\ Training\ Sample\ (B)\big)\ *\ ln(A/B).$$

3.4. Other methods and models

The prediction of the insurance field is huge and rich with many approaches and methods that are effective for forecasting the probability of claims [5–9]. Some methods cover not only the same objective as the current study but are also applied to handling more financially oriented data, missing data, and combining results of the several models [5–9]. Let's make a brief overview of these methods and present results in a general table Table 1.

Table 1
Comparison of different methods used for the insurance field

Article & year	Purpose algorithms	Algorithms	Performance metrics	The Best Model
(Smith et al. 2000) [9]	Classification to Predict Customer Retention Patterns	Decision tree (DT), Neural Networks	Accuracy ROC	Neural Networks (NN)
(Günther et al. 2014) [10]	Classification to predict the risk of leaving	Logistic regression and GAMS	ROC	Logistic regression
(Weerasinghe and Wijegunasekara 2016) [11]	Classification to predict the number of claims (low, fair, or high)	LR, DT, NN	Precision Recall Specificity	Neural networks
(Fang et al. 2016) [12]	Regression to forecast insurance customer profitability	Random Forest (RF), LR,DT Support Vector Machines (SVM), Gradient Boosting (GB)	R-squares RMSE	Random Forest
(Subudhi and Panigrahi 2017) [13]	Classification to predict insurance fraud	Decision trees, SVM, Multilayer Perceptron (MLP)	Sensitivity Specificity Accuracy	SVM
(Mau et al. 2018) [14]	Classification to predict churn, retention, and cross-selling	Random Forest	Accuracy AUC ROC F-score	RF
(Jing et al. 2018) [15]	Classification to predict claims occurrence	Naive Bayes, Bayesian, Network	Accuracy	Both have the same accuracy
(Kowshalya and Nandhini 2018) [16]	Classification to predict insurance fraud and percentage of premium amount	J48, RF, Naive Bayes	Accuracy Precision Recall	Random Forest
(Sabbeh 2018) [17]	Classification to predict churn problem	RF, AdaBoost, MLP, Stochastic GB, SVM, K-nearest Neighbor (KNN), DT, Naive Bayes, LR, Linear Discriminant Analysis (LDA)	Accuracy	AdaBoost
(Stucki 2019) [18]	Classification to predict churn and retention	LR, RF, KNN, Ada Boosting Trees, NN	Accuracy F-Score AUC	Random Forest
(Dewi et al. 2019) [19]	Regression to predict claims severity	Random forest	MSE	Random Forest
(Pesantez-Narvaez et al. 2019) [20]	Classification to predict claims occurrence	XGBoost, Logistic regression	Sensitivity Specificity Accuracy RMSE ROC	XGBoost
(Abdelhadi et al. 2020) [21]	Classification to predict claims occurrence	J48, NN, XGBoost, Naive Bayes	Accuracy ROC	XGBoost

3.4.1. Decision Tree

A decision tree (DT) and its variation is a family of classification methods that are built on a tree structure for handling the decision-making process based on binary decisions on each step. This allows to apply of the method to data with non-linear relations between features and target variables. There are several extensions of the basic model: random forest, CART models as part of multivariable trees. An example of research is in the work [22].

The random tree is used for classification tasks so a direct comparison of this method with the regression family of methods doesn't seem to be direct. There are tasks like determination of whether the claim will happen at all which can be approached by both methods but with prediction of continuous variable only one method could be used.

The simplest model is straightforward: each node checks features and directs the pipeline to one of two possible branches till the final is reached. However, this model is not suitable for complex data since it tends to overfit and variable selection can be biased.

One of the very popular extensions that was also covered by work [22] is CART or Classification and Regression Trees. It overcomes the limitation of the original model by allowing to model and predict regression variables without restricting original capabilities for categorical methods.

Another method is Random Forest. It combines several decision trees which in turn can be regressive together and via weighting of their output comes up with a single decision. It can be seen as a statistical-machine learning algorithm.

A further development that might not be so widespread in the Insurance topic but noteworthy is multivariable trees which use multivariable values for response variables.

3.4.2. Machine Learning

Support Vector Machines (SVM) is a method dedicated to providing solutions for both classification and regression problems. It is a supervised learning algorithm in which the idea is based on a hyperplane. This hyperplane of space of fewer dimensions is target one and is used for decision-making and boundary creation for target point separation.

One of the SVM key features is that in cases when it is not possible to find a plane in the current domain it can transfer inputs to higher dimensions in order to find a hyperplane in a new, higher dimension. This allows us to overcome obstacles that the target dimension possesses. SVM can be modified to solve regression tasks in [23].

4. Dataset

Necessary data for model creation were obtained from a Car Insurance dataset provided on a Kaggle web site [24]. The mentioned dataset is oriented on technical aspects of the car with most variables featuring physical parts of a machine for which police is formed.

Dataset consists of two parts which were used for training and testing. It contained 58592 and 39063 records for each part respectively. Each record is a unique policy with information about the owner of the policy and the car. The dataset has information about whether there was a claim during the upcoming 6 months for the insurance. This was a target value during the first stage of modeling. Additionally, the dataset had information about a range of different features with the total number of variables being equal to 44.

For further research, the grouping by several variables has been made with the aim of forming groups of special clients and policies for which modeling was made.

It's important to note that the dataset doesn't contain information about financial data. There is no information about the price of cars and insurance premiums for a policy.

5. Modeling Results

Modeling and forecasting have been done using generalized linear models for binomial, gaussian, and Poisson types. Scorecard was also generated to assist in decision-making and interpretation of the results.

Modeling for the probability of claim was done by building two models – binomial and Gaussian. The comparison presented in Table 2 has shown that Gaussian performs significantly better.

Table 2
Models' comparison

Model	Residuals	AIC
Binomial	3475.5	828.34
Gaussian	27300	27364

From 44 variables several were selected based on correlogram and common sense:

- ✓ age_of_car – how old is the car;
- ✓ policy_tenure – the length of the policy up to date;
- ✓ area_cluster – the area where most driving by the policy holder is done;
- ✓ make – the car's manufacturer;
- ✓ atr – synthesized variable based on the car's features: extra safety bags, lamps, etc.
- ✓ ncap_rating – rating the car's safety given by the agency.

The target relationship is then represented by the following formula:

$$is\_claim = g(k_0 + k_1 \times age\_of\_car + k_2 \times policy\_tenure + k_3 \times area\_cluster + k_4 \times make + k_5 \times atr + k_6 \times ncap\_rating).$$

It can be seen that no significant outliers in the data by judging of the distribution of the predicted values. The maximum claim probability for the whole dataset according to the model is not bigger than 0.2. This can be interpreted as uncertainty in the provided data. There are examples of claims availability and absence for the records with match all key features. All together it undermines the meaning of concentrating on one record.

The confusion matrix further highlighted the problem of such an approach. With a threshold of 0.1 it was apparent that models underperform (binomial) which is presented in Table 3 and the confusion matrix for the normal distribution which is presented in Table 4.

Table 3
Confusion matrix (binomial)

Actual \ predictions	0	1
0	50039	4805
1	3161	587

Table 4
Confusion matrix (normal)

Actual \ predictions	0	1
0	51791	3053
1	3371	377

In the next stage the modelling was made based on survival theory. It is possible to construct a survival model where each claim is treated as the death of a member of the population. We will count the length of the policy as a measure of time. Thus, the claims population "survives" during the policy length interval. It was decided that high-quality prognoses cannot be derived from existing data when claims prediction is done in the scope of the simple policy.

Let's build a Cox proportional hazards model:

$$coxph(formula = Surv(policy\_tenure, is\_claim) \sim F\ (age_{of_{policyholder}} + area_{cluster} + population\_density, data = car\_insurance\_tibble)),$$

where $n = 58592$, number of events $= 3748$.

It can be seen however that length of policy indeed has an effect on the claims number amounts but this observation is rather trivial and cannot be used to make a decision since only short-range policies should be preferred (Figure 1). Therefore, the relationship between the length of the policy and the frequency of lawsuits was revealed. At the moment of time 1, 1.6 and 1.7 year duration there is a sharp increase in claims. It is possible to perform separation and in the future to focus on the threshold values found. Also from the survival model is easier to determine the duration of the most risky policies and to define the possible new policies politics.

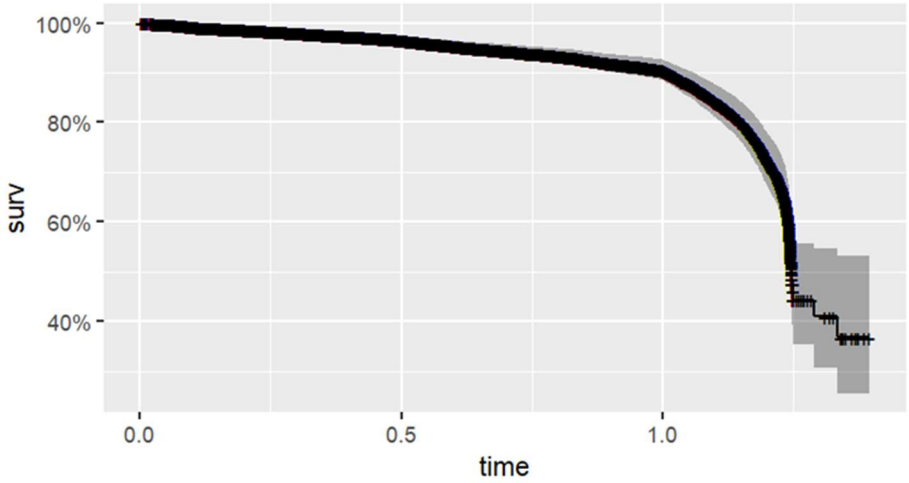

Figure 1: Survival model for the insurance policies

The calculation of individual cases (a claim for each policy separately) showed the absence of parameters and characteristics that would accurately indicate the onset of a claim. All probabilities for each policy lie between 0.001 and 0.12. In this case, a decision was made to proceed to the consideration of individual segments.

Grouping of data by segment, manufacturer, and machine brand was performed. Thus, we have moved from looking at an individual car to the segment as a whole, where individual characteristics are of little importance.

Two values can be calculated for segments:

1. The number of claims in the segment.
2. Amount of payment by segments.

It was decided to implement a further approach to working with groups. The Poisson generalized linear model was chosen as the model to forecast the number of cases. It showed a high level of accuracy.

The equation for modelling relationships was presented in the such way:

$$evaluated\_N = g(total + segment + area\_cluster + airbags + \\ + make + policy\_tenure + age\_of\_car).$$

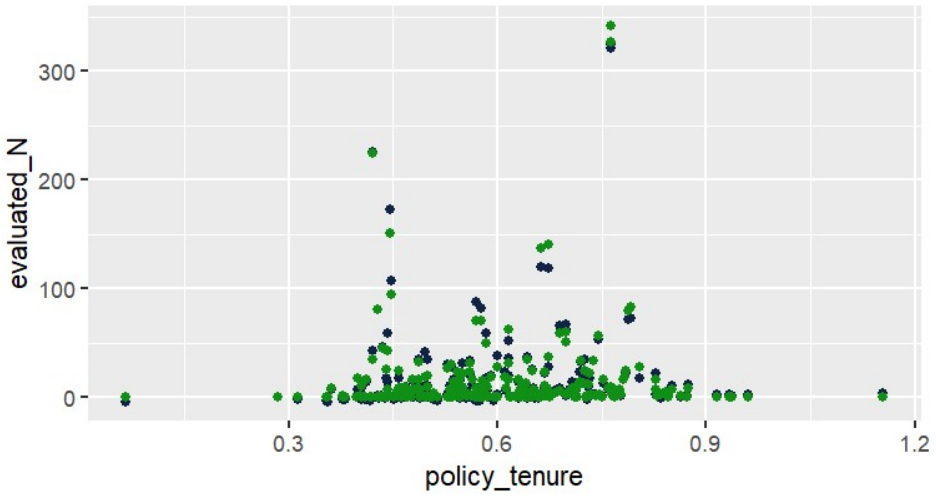

Figure 2: Real (black) and estimated (green) values plotted together

Table 5

Model	Residuals	AIC
Poisson	629.27	1425.6
Normal	1425.6	5952.5

As can be seen in Figures 3 and 4 the claims' number prediction across groups has a higher quality degree. This also shows that despite the low ability to predict each unique case, prediction of the group is a much easier task.

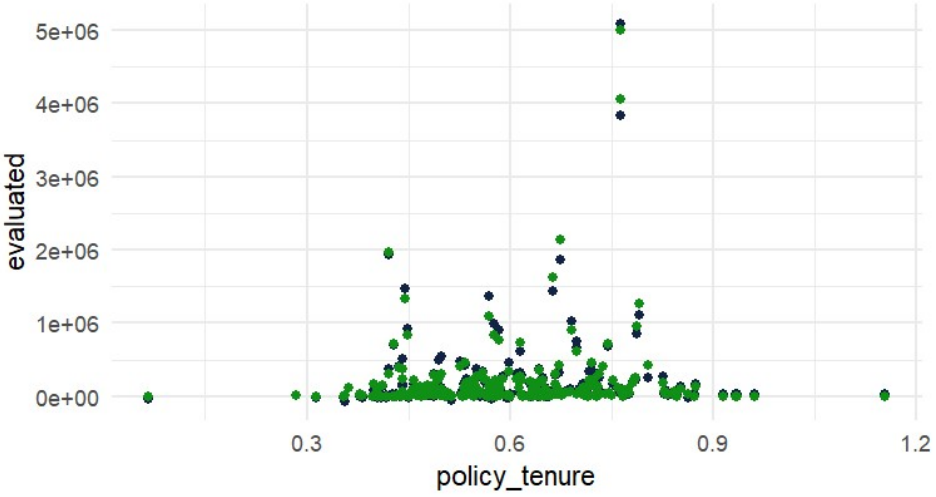

Figure 3: Value of claimed cars (black) and estimated (green)

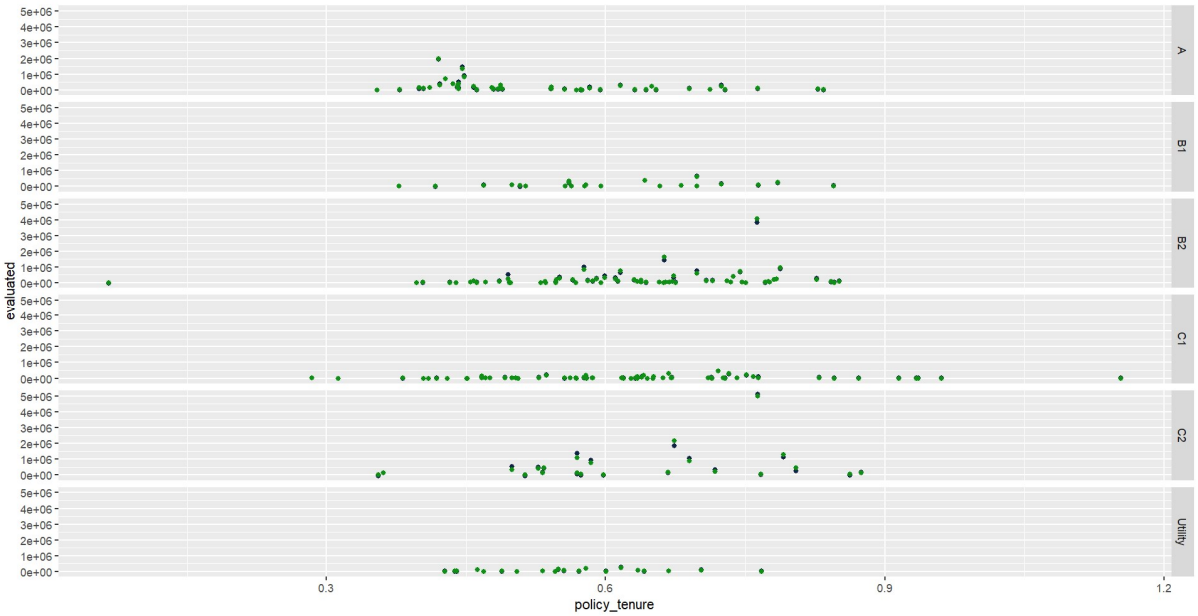

Figure 4: Against segments to which cars belong A-Utility

Another approach was chosen for dealing with the group. It was about forecasting the price of all cars for which claims were issued. The gaussian model was used as the most appropriate. This also showed significant accuracy (Table 6).

Additionally, in the dataset, the car's price was missing data. For this model, the following approach was used:

1. To find the average price for every class.
2. Adjust it according to the attribute feature.
3. To group price per category to create a new feature – total (price).

The equation for modeling is as follows:

$$paid\_price = g(total + group\_price + segment + policy\_tenure + \\ + age\_of\_car).$$

Table 6
Result

Model	Residuals	AIC
Normal	8.4797e+11	5886.3

We need to understand which variables and intervals for these variables are the most significant in the aim of our insurance task. Information value (IV) is one of the most useful techniques for selecting important variables in a predictive model. This helps to rank the variables based on their importance. On Figure 5 it is presented how many claims cases were and how they correlated in accordance to different values of the car's age.

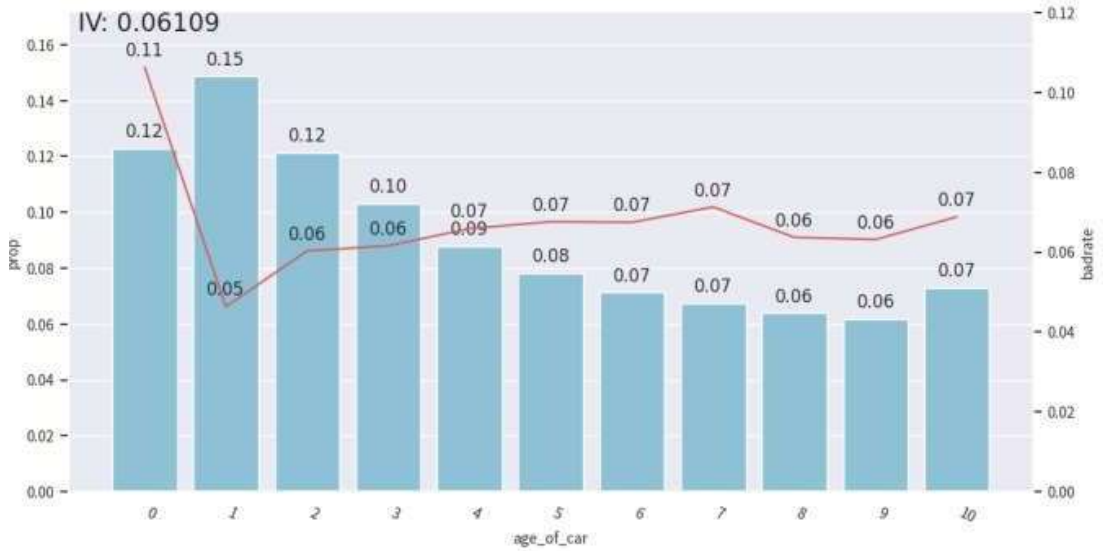

Figure 5. Informational value of the variable age_of_car

Finally, the scorecard was built (Table 7). It provided information about values that are associated with high risk of a claim for this dataset. Non-significant values have been filtered out. The remaining variables describe continuous data – age of policyholder and policy tenure for which binning is made. Categorical variables were also presented in the work – area of clusters which were named in the initial dataset and ranges from C1 to C22 and variables that related to technical aspects: rear mirror availability and functionality, brakes type, and transmission type.

Table 7
Scorecard for a claim's prediction

Number of interval	Variable	Binning	Score
0	age_of_policyholder	[-inf ~ 0.384615384615385)	0.74
1	age_of_policyholder	[0.384615384615385 ~ 0.442307692307692)	0.06
2	age_of_policyholder	[0.442307692307692 ~ 0.490384615384615)	-0.51
3	age_of_policyholder	[0.490384615384615 ~ 0.634615384615385)	0.25
4	age_of_policyholder	[0.634615384615385 ~ inf)	-0.84
0	area_cluster	C17,C20,C9,C7,C1,C10,C15	2.61
1	area_cluster	C16,C13,C5,C12,C6	1.13
2	area_cluster	C11,C3,C2,C8	-0.79
3	area_cluster	C4,C19,C14,C22,C21,C18	-2.13
0	policy_tenure	[-inf ~ 0.211309751692924)	5.3
1	policy_tenure	[0.211309751692924 ~ 0.813392835491761)	1.49
2	policy_tenure	[0.813392835491761 ~ inf)	-3.86
0	is_day_night_rear_view_mirror	No	0
1	is_day_night_rear_view_mirror	Yes	0.26
0	steering_type	Manual,Power	0.05
1	steering_type	Electric	0.17
0	rear_brakes_type	Drum	0.05
1	rear_brakes_type	Disc	0.26
0	is_tpms	No	0.05
1	is_tpms	Yes	0.26
0	make	[-inf ~ 2)	0.05
1	make	[2 ~ inf)	0.19
0	transmission_type	Manual	0.05

1	transmission_type	Automatic	0.2
0	is_rear_window_washer	No	0.08
1	is_rear_window_washer	Yes	0.14
0	is_rear_window_wiper	No	0.08
1	is_rear_window_wiper	Yes	0.14
0	is_rear_window_defogger	No	0.15
1	is_rear_window_defogger	Yes	-0.03

6. Conclusion

Today car insurance companies require a lot of information to decide policies and conditions [5]. Even though a vast amount of information can be collected it doesn't guarantee the ability to create a model that can predict a claim for a specific policy with a significant level of accuracy due to the randomness the of claim's nature. Some special cases can be chosen, less or more prone to claims cases can be selected but it doesn't allow to make a robust prediction according to the results. From a built model for probability prediction, the gaussian generalized model has been chosen. It shows that claims' nature cannot be determined based on some specific features or its combinations since for same key variables. There are examples of policies with and without claims. Obtained values show a high level of centering which doesn't allow to select intervals for confident claim selection and hence undermines the usefulness of such an approach.

The problem of single-claim prediction is the hardest one. For the claims risk management, we need to forecast the probability of each claim, of each type of claim, and to develop a special scoring card in an understandable and easily interpretable manner with the key features automatically.

More promising are results for a group of claims where policies are selected and combined under the same group with similar features. Such groups have a higher degree of an accuracy and can be modeled and forecasted with respect to number of claims or total cars' price for which claims have been made. Overall, the results show a low ability to predict specific cases but relatively high confidence in forecasting in big groups.

It is worth noting that different methods like Random Forest could perform better with the task of predicting claims per observation which can be examined in consequent researchers.

Finally, the scorecard is a high-quality tool to make decisions for clients directly. It is not only easy to interpret but to use. We used the scorecard to determine in an understandable and easily interpretable manner the key features. It yields great results on the grouped data and provides valuable insights about the tendencies. It is also useful to implement the scorecards instrument as a good tool for telecom and different finance for the big data tasks [25, 26] where it is needed to evaluate some scores and influence of characteristics as well.

References

[1] Wang, Xin, Tapani Ahonen, and Jari Nurmi. "Applying CDMA technique to network-on- M. Denuit, X. Marechal, S. Pitrebois, J. Walhin, Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems, Wiley, 2007.

[2] R. Verbelen, K. Antonio, G. Claeskens, "Unravelling the predictive power of telematics data in car insurance pricing", Journal of the Royal Statistical Society, Series C (Applied Statistics), 67.5 (2018): 1275-1304. doi:10.1111/rssc.12283.

[3] J. Ashworth Nelder, R. W.M. Wedderburn, "Generalized linear models", Journal of the Royal Statistical Society: Series A (General), 135.3 (1972): 370-384. doi:10.2307/2344614.

[4] N. V. Kuznietsova, P. I. Bidyuk, Theory and practice of financial risk analysis: systemic approach, Lira-K, Kyiv, 2020.

[5] V. Selvakumar, D. K. Satpathi, P. T. V. Praveen Kumar, V. V. Haragopal, Predictive Modeling of Insurance Claims Using Machine Learning Approach for Different Types of Motor Vehicles, Universal Journal of Accounting and Finance 9(1): 1-14, 2021. doi: 10.13189/ujaf.2021.090101.
[6] C. Ye, L. Zhang, M. Han, Y. Yu, B. Zhao, Y. Yang, Combining Predictions of Auto Insurance Claims, Econometrics 2022, 10(2), 19. doi: 10.3390/econometrics10020019.
[7] W. Yu, G. Guan, J. Li, Q. Wang, X. Xie, Y. Zhang, Y. Huang, X. Yu, C. Cui, Claim Amount Forecasting and Pricing of Automobile Insurance Based on the BP Neural Network, Hindawi Complexity Volume 2021. doi: 10.1155/2021/6616121.
[8] M. Hanafy, R. Ming, Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 6, 2021. doi: 10.14569/IJACSA.2021.0120656.
[9] K. A. Smith, R. J. Willis, M. Brooks, An analysis of customer retention and insurance claim patterns using data mining: a case study, Journal of the Operational Research Society 51: 532–41. doi: 10.1057/palgrave.jors.2600941.
[10] C.-C. Günther, I. F. Tvete , K. Aas, G. I. Sandnes, Ø. Borgan, Modelling and predicting customer churn from an insurance company, Scandinavian Actuarial Journal 2014 Vol. 2014, No. 1, 58–71. doi: 10.1080/03461238.2011.636502.
[11] K.P.M.L.P. Weerasinghe, M.C. Wijegunasekara, A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims, European International Journal of Science and Technology Vol. 5 No. 1 January, 2016.
[12] K. Fang, Y. Jiang, M. Song, Customer profitability forecasting using Big Data Analytics: A case study of the insurance industry, Computers & Industrial Engineering (2016). doi: 10.1016/ j.cie.2016.09.011.
[13] S. Subudhi, S. Panigrahi, Use of Optimized Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection, Journal of King Saud University – Computer and Information Sciences (2017). doi: 10.1016/j.jksuci.2017.09.010.
[14] S. Mau, I. Pletikosa, J. Wagner, Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments, International Journal of Bank Marketing 36 (2018): 6. doi: 10.1108/IJBM-11-2016-0180.
[15] L. Jing, W. Zhao, K. Sharma, R. Feng, Research on Probability-based Learning Application on Car Insurance Data, proceedings of the 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017), Amsterdam: Atlantis Press, 2018. doi: 10.2991/macmc-17.2018.14.
[16] G. Kowshalya, M. Nandhini, Predicting fraudulent claims in automobile insurance, in: Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, April 20–21; pp. 1338–43. doi: 10.1109/ICICCT.2018.8473034.
[17] S. F. Sabbeh, Machine-learning techniques for customer retention: A comparative study, International Journal of Advanced Computer Science and Applications Vol. 9, No. 2, 2018: 273–81.
[18] O. Stucki, Predicting the Customer Churn with Machine Learning Methods: Case: Private Insurance Customer Data, Master's dissertation, LUT University, Lappeenranta, Finland, 2019.
[19] K. C. Dewi, H. Murfi, S. Abdullah, Analysis Accuracy of Random Forest Model for Big Data – A Case Study of Claim Severity Prediction in Car Insurance, in: Paper presented at 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, October 23–24; pp. 60–65. doi: 10.1109/ICSITech46713.2019.8987520.
[20] J. Pesantez-Narvaez, M. Guillen, M. Alcañiz, Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression, Risks 2019, 7(2), 70. doi: 10.3390/risks7020070.
[21] S. Abdelhadi, K. Elbahnasy, M. Abdelsalam, A proposed model to predict auto insurance claims using machine learning techniques, Journal of Theoretical and Applied Information Technology 30th November 2020. Vol.98. No 22: 3428–3437.

[22] Z. Quan, E. A. Valdez, Predictive analytics of insurance claims using multivaria te decision trees, Depend. Model. 2018; 6:377–407. doi: 10.1515/demo-2018-0022.
[23] E. Alamir, T. Urgessa, A. Hunegnaw , T. Gopikrishna, Motor Insurance Claim Status Prediction using Machine Learning Techniques, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 3, 2021. doi: 10.14569/IJACSA.2021.0120354.
[24] Kaggle. URL: https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification?resource=download
[25] Kuznietsova, N., Bidyuk, P., Kuznietsova, M. Data Mining Methods, Models and Solutions for Big Data Cases in Telecommunication Industry, Lecture Notes on Data Engineering and Communications Technologiesthis link is disabled, 2022, 77, pp. 107–127. doi: 10.1007/978-3-030-82014-5_8.
[26] Kuznietsova N., Bateiko E. Analysis and Development of Mathematical Models for Assessing Investment Risks in Financial Markets. CEUR Workshop Proceeding (ISSN 1613-0073). 2023. Vol. 3503, p. 92-101. https://ceur-ws.org/Vol-3503/paper9.pdf.