ORIGINAL ARTICLE
DOI: 10.5433/1679-0375.2020v41n1p79

Vehicle claims in the south of Minas Gerais: an approach using classification models

Ocorrência de sinistros em veículos no sul de Minas Gerais: uma abordagem via modelos de classificação

Luiz Otávio de Oliveira Pala[1]; Marcela de Marillac Carvalho[2];
Paulo Henrique Sales Guimarães[3]; Thelma Sáfadi[4]

Abstract

With the changes in the patterns of risk, new insurance products are available on the market. Consequently, pricing models are restructured to manage levels of risk and create premiums that maintain the well-being of insurers. This work analyzed the Logistics and Random forests models in the classification of total loss events in the south of Minas Gerais using original and artificial samples, built by the ROSE resampling method, which is a procedure for constructing artificial samples in a smoothing bootstrap. A total loss of a vehicle is considered when the repair costs for the same event exceed a percentage established by contract. As a result, it was obtained that the models with artificial data improved the balanced accuracy rate on unbalanced data.

Keywords: Random forest. Random over sampling examples. Logistic regression.

Resumo

Com as mudanças nos padrões de risco, novos produtos de seguros são disponibilizados no mercado, atendendo as demandas do consumidor. Consequentemente, os modelos de precificação são reestruturados de modo a gerenciar os níveis de risco e estabelecer prêmios que mantenham o bem estar atuarial, alocando apólices em carteiras através de modelos de classificação e *clusterização*. Este trabalho analisou o desempenho dos modelos Logísticos e *Random forests* na classificação de ocorrências de sinistros do tipo colisão por perda total no sul de Minas Gerais utilizando amostras de treino originais e artificiais via método de reamostragem ROSE, que é um procedimento de construção de amostras artificiais em uma suavização *bootstrap*. Considera-se a perda total de um veículo quando os custos de reparos do sinistro de um mesmo evento superarem um percentual estabelecido contratualmente. Como resultado, obteve-se que os modelos com amostra artificial apresentaram resultados de acurácia balanceada superiores aos demais, indicando a melhoria através de métodos de reamostragem durante o treino.

Palavras-chave: Random forest. Random over sampling examples. Regressão logística.

[1] Doutorando no Prog. de Estatística e Exp. Agropecuária, UFLA, Lavras, MG, Brasil; E-mail: luiz.pala@estudante.ufla.br
[2] Doutoranda no Prog. de Estatística e Exp. Agropecuária, UFLA, Lavras, MG, Brasil; E-mail: marcela.carvalho2@estudante.ufla.br
[3] Prof. Dr., Depto. de Estatística, UFLA, Lavras, MG, Brasil; E-mail: paulo.guimaraes@des.ufla.br
[4] Profa. Dr[a]., Depto. de Estatística, UFLA, Lavras, MG, Brasil; E-mail: safadi@des.ufla.br

Introduction

In the historical context of the actuarial field, the development of aversion and prevention forms to conditions that influence or compromise an individual's life and assets in the form of random events can be noticed (FILHO, 2011). The emergence of insurance, for example, was constituted as a way of repairing damages caused by the occurrence of a certain risk, priced in an equivalence relation between the premium charged and the risk assumed by the insurance company (ZEMIACKI, 2006).

In the scope of pricing Blake *et al.* (2013) and Dionne (2013) indicate that demographic and economic problems have an impact on the insurers provisions, making the risk management methods, portfolio segmentation and precision in premium estimates increasingly necessary.

The improvement of premium calculations has been playing an important role in insurance companies, making them seek better ways of classifying customers, commonly from generalized linear models and machine learning models (SPEDICATO; DUTANG; PETRINI, 2018). Classifying methods are being used in areas such as finance and epidemiology, identifying fraud and disease, respectively (MENARDI; TORELLI, 2014) and in industrial areas, such as analysis of biodiesel samples (FILHO *et al*., 2015). Note that this process can be expanded to the insurance market in the classification of clients with higher risk, as seen in Spedicato, Dutang and Petrini (2018).

In the automobile industry, these methods can be used by grouping policies with a higher risk of total loss or fire, for example. The total loss of a vehicle is defined when the costs resulting from the same event exceed a percentage. This percentage is set in value determined coverage policies, up to 75% of the value determined and market value coverage, up to 75% of the value of the vehicle calculated via the adjustment factor, established by the Circular Letter No. 145, Article 16 (SUSEP, 2000).

In the machine learning methods, the decision trees and random forests methods have been used in classification phenomena. By definition, random forest is a classifier made up of a set of decision trees $\{h(x,\Theta_k), k = 1, ...\}$ since $\{\Theta_k\}$ are independent and identically distributed random vectors (BREIMAN, 2001). This classifier is shown as a versatile and simple machine learning approach (LANTZ, 2015).

According to Hastie, Tibshrani and Friedman (2008), in the classification scenario the random forest model is built from B bootstrap samples Z^* of the training set. For each of these samples, a T_b tree will be formed from the selection of m (typically m = $\sqrt{p}$) among the total of p variables. Since $\hat{C}_b(x)$ is the prediction of the j-th tree, the prediction will be the result of the majority of votes from the trees.

The process of building a tree is based on creating mutually exclusive rectangles as follows: i) divide the space of the covariates $\mathbf{X}$ into M regions $(R_1, R_2, ...)$ built to reduce the classification error rate; ii) for each y_i belonging to R_j region, it is taken the mode of y as predictor. According to the number of possible partitions, the binary recursive division algorithm is commonly used (MORETTIN; SINGER, 2020). More detail on classification trees and random forests are in Hastie, Tibshrani and Friedman (2008), Lantz (2015), Izbicki and Santos (2019) and Morettin and Singer (2020).

The logistic regression model can also be used for the classification. This is commonly applied in phenomena where the dependent variable is categorical (SUSAC *et al*., 2016). Hence, the response variable classification happens establishing a threshold in the predictions of the logistic model.

One of the problems discussed in the literature is the quality of classifiers in unbalanced scenario and the presence of rare events (MENARDI; TORELLI, 2014; PIERRI; STANGHELLINI; BISTONI, 2016; LIN *et al*., 2017). One of the methods used in binary classification problems is Random Over-Sampling Examples (ROSE), as seen in Tantithamthavorn *et al*. (2018) and Zhang and Chen (2019).

The aim of this work is to analyze the occurrence of collision total loss claims of classification models under unbalanced scenarios of the data from the south of Minas Gerais. For this purpose, four classification models were used, random forests and logistics, considering original and artificial training samples of the ROSE method to compare predictive power. The models were built on R language (R CORE TEAM, 2020).

Materials and methods

The data used in this study were provided by the Superintendência de Seguros Privados (SUSEP), regarding 73,702 policies effective for at least one day in the first semester of 2019 in all tariff categories.

The missing observations was from the range, representing 0.01085% of total lines. The base considers diverse covariates, of which five was selected by the study, for instance:

- Owner: [1]: related to forms of hiring. Distributed as follows:
  i) natural person, male (M);
  ii) natural person, female (F);
  iii) juridical person (J);
  iv) no information (0).
- Exposure: time of exposure of each policy in the observed period, being the best estimator of the number of insured vehicles (SUSEP, 2020);
- Vehicle year: variable that corresponds to the year of the insured vehicle;
- Age: categorical variable on in the levels:
  i) not informed age (0);
  ii) between 18 and 25 years old (1);
  iii) between 26 and 35 years old (2);
  iv) between 36 and 45 years old (3);
  v) between 46 and 55 years old (4);
  vi) more than 55 years old (5).
- Frequency: number of total collision coverage claims.

The Occurrence and Year variables were created based on the Vehicle year and Frequency variables. These variables were specified as follows: Occurrence is a binary variable that assumes unit value if the Frequency is higher than 1 as well as the Year variable, assuming a value of 1 if the year of the vehicle is equal to or higher than the year of 2010.

A descriptive analysis based on the mentioned variables was made, followed by the random forests model implementation with 100 trees and $\sqrt{p}$ [2] applicants covariates in each partition, in order to predict the occurrence, or not, of the claims.

The logistic models were implemented in a similar way, in order that p assumes the probability of occurrence of an event, the logistical function is expressed according to the equation (1)

$$p = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_r x_r}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_r x_r}}, \quad (1)$$

denoting g(x) = $\beta_0+\beta_1 x_1+\ldots+\beta_r x_r$, a linear relationship correspondent to the equation (2)

$$logit(y) = ln\left\{\frac{p}{1-p}\right\} = ln\left\{\frac{\frac{e^{g(x)}}{1+e^{g(x)}}}{1-\frac{e^{g(x)}}{1+e^{g(x)}}}\right\}, \quad (2)$$

being x_i, i= 1,..., r, the covariables, y the response variable and β_j, j = 1,.., r the estimated coefficients (SUSAC *et al.*, 2016; SARLIJA; BILANDZIC; STANIC, 2017). Thus, the response variable classification was obtained establishing a threshold of p =0.5, which is commonly adopted by literature, as seen in Yao *et al.* (2019). In practice, this threshold can be established according to the insurer's risk policy.

The process of cross-validation in 10-folds was used for modeling, dividing the samples into training (90%) and trial (10%) in fixed source for the generation of pseudorandom numbers. According to Hastie *et al.* (2008), there is no general rule for establishing the training and test percentages. However, studies usually adhere percentages between 75% and 90% for modeling.

Due to the unbalance which is the presence of a rare class, occurrence of the claim, four models were built. Two of them were built from the original set (T_n) and the rest from the artificial set ($T_n{}^*$) generated by the Randomly Over Sampling Examples (ROSE) method, which is used in unbalanced scenarios.

According to Menardi and Torelli (2014), the ROSE method aims to build artificial samples in a smooth bootstrap. Defining $\mathbf{T}_n$ as a set of data, y as a response in classes $\mathscr{Y} \in \{y_0, y_1\}$, a set of covariates X and n_j the size of $\mathscr{Y}_j$, j = 0, 1. Therefore:

i) Select $y = \mathscr{Y}_j \in \mathbf{T}_n$ with probability $\frac{1}{2}$;

ii) Selecting $(x_i, y_i) \in \mathbf{T}_n$ so that $y_i = y$ has probability $p_i = \frac{1}{n_j}$;

iii) Sample X of a kernel $KH_j(.,x_i)$ since KH_j is a distribution centered on x_i and H_j is the matrix from the scales parameters, optimized assuming a Gaussian kernel. In sum, the method generates new observations, treated as artificial samples, in the neighborhood of each one of the classes, where the size of this proximity is weighted by H_j.

[1] Variable treated as gender on the Superintendência de Seguros Privados platform.

[2] Since p is the number of present covariables.

Subsequently, the quality metrics of classification models such as accuracy rates, balanced accuracy and Kappa (k) index were evaluated in confusion matrices. Since $M_{2,2}$ is a confusion matrix with the average percentage of classification according to cross-validation, k can be obtained as follows:

$$\mathrm{k} = \frac{P_0 - P_c}{1 - P_c}, \tag{3}$$

being $P_0 = Tr(M)$ and $P_c = \sum_{i=1}^{2} m_{1i} \sum_{i=1}^{2} m_{i1} + \sum_{i=1}^{2} m_{2i} \sum_{i=1}^{2} m_{i2}$.

This index assesses the agreement between the expected and observed classification, in order to the higher the value of k ($\mathrm{k} \to 1$), the better agreement of the model. In the confusion matrix, the accuracy (AC) is equal to P_0, and the balanced accuracy (ACB) is given by:

$$\mathrm{ACB} = \frac{1}{2}\{m_{11}(m_{11}+m_{21})^{-1} + m_{22}(m_{12}+m_{22})^{-1}\}. \tag{4}$$

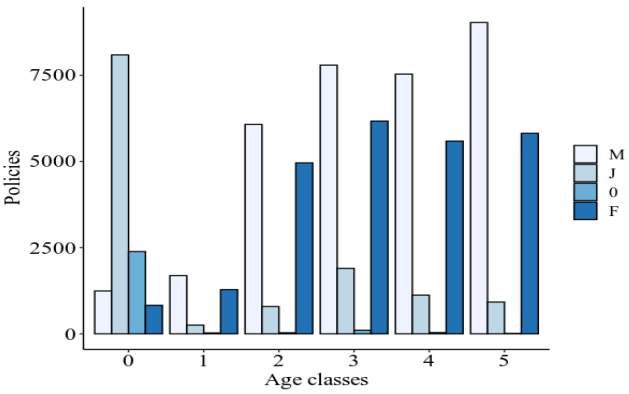
In addition to the accuracy results, the following Receiver Operating Characteristic Curve (ROC) models were built to compare the prediction performance for each of the 10-folds. This curve is used in binary classification problems, evaluating the performance of the model in a graphical representation of trade-off between the false positive and true positive type classification rates (MCCLISH, 1989), therefore curves that are more distant from the main diagonal depict a better model performance to the domain (PRATI; BATISTA; MONARD, 2008).

For the analysis, the ROSE packages (LUNARDON; MENARDI; TORELLI, 2015) the randomForest (LIAW; WIENER, 2018), available in R (R CORE TEAM, 2020) were used.

Results and discussion

Considering the 73,694 structured observations, it can be noticed in Figure 1, the age class distribution in relation to the class of owners. Note that in the age class without information (0) there is a predominance of policies from legal entities (J), corresponding to 8,091 policies, a fact that is expected. There is a higher number of male insured policies (M) in the other groups, mainly in the groups older than 55 years old. In the female group (F), the highest concentration is related to the class 3, that is, women aged between 36 and 45 years old. It should be noted that the classes represent approximately 17.03%, 4.41%, 16.10%, 21.67%, 19.38% and 21.41%, respectively for groups from 0 to 5.

Figure 1 – Distribution of the number of policies in relation to age classes and owners



Source: The authors.

Regarding the 865 claims that occurred in the first semester of 2019, it can be seen in Table 1 that the higher percentage of these was due to the female classes or male (33.4% and 53.8%, respectively). Regarding age classes, categories 3 and 5 were the most representative, that is, individuals aged between 36 and 45 years old, or over 55 years old.

Table 1 – Percentage of 865 claims in the first semester of 2019 according to classes age and owners (P.)

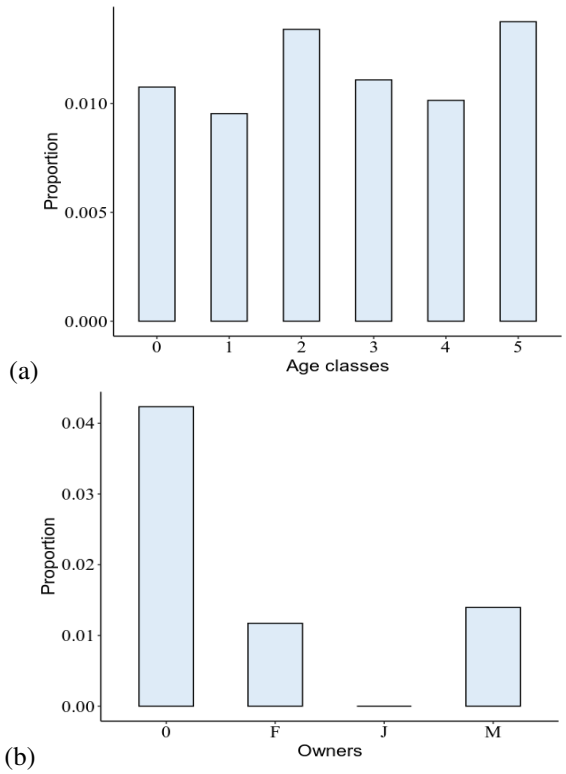
	Age classes						
P.	0	1	2	3	4	5	∑
0	11.8	0.00	0.12	0.69	0.12	0.00	12.7
F	1.16	1.27	7.05	7.40	8.21	8.32	33.4
J	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M	2.66	2.31	11.3	12.37	8.44	16.76	53.8
∑	15.6	3.60	18.5	20.5	16.8	25.1	100

Source: The authors.

However, there may be an effect on the number of policies in each class, as it can be seen in Figure 1. For example, there are 33,364 policies from male owners and 2,598 of the class without information about the owner. So, more claims are expected to occur in a class with a higher number of policies in relation to the class without information. Thus, the occurrences were standardized according to the number of policies in each class, shown in Figure 2.

In relation to the proportion of claims adjusted by the number of policies in each age category, the classes 2 and 5 had the highest proportions. In other words, individuals aged from 26 to 35 years old and over 55 years, Figure 2a. Considering the classes of owners, the highest adjusted percentage came from policies without information about the owners, and the legal level showed the lowest percentage, a fact that can be seen in the Figure 2b.

Figure 2 – Proportion of claims occurrence in each age category (a) and each category of owners (b) adjusted for the number of policies in the age classes and owners



Source: The authors.

After a brief exploratory analysis, the artificial sample was created from the original set and the process of cross-validation was specified at k = 10-folds. In T_n, the variable Occurrence showed 865 unit results, representing 1.173%.

In contrast, the set T_n^* was generated with the probability of the minority class $p = 0.5$. From that, there is a higher balance between classes, with unitary results at 50.16% of T_n^*. The two sets were used on random forests and logistics models for the variable response classification. The model considered was defined as follows:

$$\text{Occurrence} \sim \text{Owner} + \text{Age} + \text{Exposition} + \text{Year}.$$

The predictive quality measures were assessed with respect models, as the accuracy, balanced accuracy and the kappa index. Note that through Table 2 the models using T_n showed improved accuracy when compared to the T_n^*. With the artificial samples application there was increasing on kappa, indicating a predictive improvement, still less than 0.8 in identifying the occurrence of claims. The same fact occurred in the balanced accuracies, that are typically used for models comparison in unbalanced scenarios.

Table 2 – Measures of average quality of the built models considering the original sample (T_n) and the artificial sample (T_n^*)

Model	Sample	AC	K	ACB
Random Forest	T_n	0.988	0.000	0.500
	T_n^*	0.806	0.612	0.806
Logistic	T_n	0.988	0.000	0.500
	T_n^*	0.738	0.476	0.738

Nota: Accuracy (AC); Kappa (K); Balanced accuracy (ACB).
Source: The authors.

Analyzing the confusion matrix available in Table 3, it can be noticed that the models with T_n^* obtained an increase in the detection of claims, however there was a cost associated with this improvement regarding to the errors in the class where the claim did not occur.

Table 3 – Percentage average classification according to the cross-validation of the models: (a) Random forest with T_n (b) Random forest with T_n^*, (c) Logistics with T_n, (d) Logistics with T_n^*

(a)

Predicted \ Occurrence	0	1
0	98.8	1.20
1	0.00	0.00

(b)

Predicted \ Occurrence	0	1
0	41.6	11.1
1	8.30	39.1

(c)

Predicted \ Occurrence	0	1
0	98.8	1.20
1	0.00	0.00

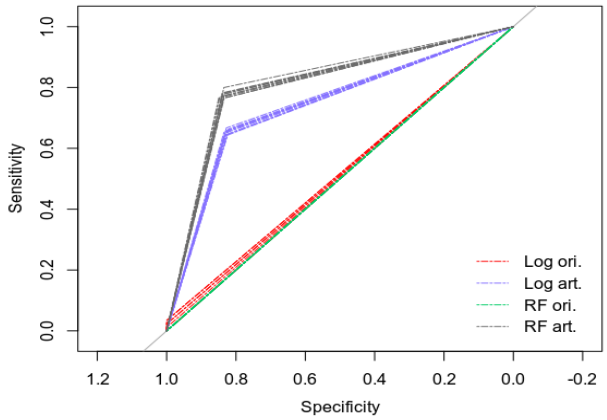
(d)

Predicted \ Occurrence	0	1
0	40.8	17.2
1	9.00	33.0

Source: The authors.

To evaluate the quality of the classification models, Figure 3 shows the behavior of the ROC curves of the models built in each of the 10-folds. Note that the random forest model with the original sample did not present a good rating performance, the same happened with the trained logistic model in the original sample. Alternatively considering the ROC curve, the best ratings were those that used the artificial sample, with bigger areas under the curve. All the others presented random rating behaviors, with low power of distinction.

Figure 3 – ROC curves of the models built in each of the 10-folds



Source: The authors.

The conditions pointed out as the increase in the area under the curve and balanced accuracy of the trained models with artificial samples show the best predictive capacity model used in balanced phenomena or without the presence of rare events. The ROSE method contributed to the detection of the occurrence of claims in logistic and random forest models, assisting in the classification of this type of event.

Final remarks

Due the unbalanced level, in other words, only a few total loss claims occurrence, the ROSE method generated improvements in the capability of classification capacity in both models, elevating the balanced accuracy and the quality of the ROC curve. However, even with the increasing on quality from artificial sample models, there was a high level of errors of false positive and negative type.

This fact generates a contrast between the use of both types of sample forms, as follows:

i) keeping a minimal of false-positive errors considering the original sample, but with low detection power around truly negatives, or

ii) elevating the false-positive errors using the artificial sample, but increasing the truly negatives detection capability. These decisions must be evaluated in order to keep the economic and financial welfare by the insurance companies.

Furthermore, it is necessary by the insurance company to constitute error thresholds of false positive and negative types, because it can affect an individual's premium when buying an insurance product.

To expand and explore the total loss type of claims in the south of Minas Gerais, the main goal is to use and compare approaches as neural networks, naive Bayes and ensemble learning classes in relation on methods used in this study, aside from studying decisions making thresholds on logistic regression for this insurances context.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

BLAKE, D.; CAIRNS, A.; COUGHLAN, G.; DOWD, K.; MACMINN, R. The new life market. *Journal of Risk and Insurance*. Orlando, v. 80, n. 3, p. 501-558, 2013. DOI: https://doi.org/10.1111/j.1539-6975.2012.01514.x.

BREIMAN, L. Random Forests. *Kluwer Academic Publishers*. United States], v. 45, p. 5-32, 2001.

DIONNE, G. Risk management: history, definition, and critique. *Risk Management and Insurance Review*. Hoboken, v. 16, n. 2, p. 147-166, 2013. DOI: http://dx.doi.org/10.2139/ssrn.2231635.

FILHO, O. *Seguros*: fundamentos, formação de preço, provisões e funções biométricas. Editora Atlas. São Paulo. 2011.

FILHO, V. M. C.; SPACINO, K. R.; ROMAGNOLI, É. S.; SILVA, L. R. C.; BORSATO, D. Perfil do biodiesel B100 comercializado na região de Londrina: aplicação de redes neurais do tipo mapa auto-organizável. *Semina*: Ciências Exatas e Tecnológicas, v. 36, n. 2, p. 63-70. 2015. DOI: http://dx.doi.org/10.5433/1679-0375.2015v36n2p63.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. 2. ed. New York: Springer Series in Statistics. 2008.

IZBICK, R.; SANTOS, T. *Machine Learning sob a ótica estatística: uma abordagem preditivista para a estatística com exemplos em R*. [São Carlos: *s. n.*], 2019. Available in: <http://www.rizbicki.ufscar.br/sml.pdf>. Access in: Nov. 2019.

LANTZ, B. *Machine learning with R*. 2. ed. [*s. l.*]: Packt Publishing, 2015.

LIAW, A.; WIENER, M. *Breiman and Cutler's Random Forests for Classification and Regression*. 2018. Available in: <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf>. Access in: Mar. 2020.

LIN, W.; WU, Z.; LIN, L.; WEN, A.; LI, J. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE*, Guangzhou, p. 531-536, 2017. DOI: https://doi.org/10.1109/CSE-EUC.2017.99.

LUNARDON, N.; MENARDI, G.; TORELLI, N. *ROSE*: Random Over-Sampling Examples. 2015. Available in: <https://cran.r-project.org/web/packages/ROSE/ROSE.pdf>. Access in: Jan. 2020.

MCCLISH, D. Analyzing a Portion of the ROC Curve. *Medical Decision Making*. v. 9, n. 3. 1989. DOI: https://doi.org/10.1177/0272989X8900900307.

MENARDI, G.; TORELLI, N. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*. Hoboken, v. 28, p. 92-122, 2014. DOI: https://doi.org/10.1007/s10618-012-0295-5.

MORETTIN, P.; SINGER, J. *Introdução à ciência de dados: fundamentos e aplicações*. São Paulo: USP, 2020. Available in: <https://www.ime.usp.br/~pam/>. Access in: Mar. 2020.

PRATI, R.; BATISTA, G.; MONARD, M.. Curvas ROC para avaliação de classificadores. *IEEE*. Guangzhou, v. 6, n. 2, 2008. Available in: <https://sites.icmc.usp.br/gbatista/files/ieee_la2008.pdf>. Access in: Jan. 2020.

PIERRI, F.; STANGHELLINI, E.; BISTONI, N. Risk analysis and retrospective unbalanced data. *REVSTAT*. Lisboa, v. 14, n. 2, p. 157-169, 2016. Available in: <https://www.ine.pt/revstat/pdf/rs160204.pdf>. Access in: Mar. 2020.

R CORE TEAM. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, Vienna, Austria. 2020.

SARLIJA, N.; BILANDZIC, A.; STANIC, M. Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models. *Croatian Operational Research Review*.Croatia, v. 8, p. 631-652, 2017. DOI: https://doi.org/10.17535/crorr.2017.0041.

SPEDICATO, G.; DUTANG, C.; PETRINI, L.. Machine Learning methods to perform pricing optimization: a comparison with Standard Generalized Linear Models. *Variance Journal*, Arlington, v. 12, n. 1, p. 69-89, 2018. Available in: <https://www.variancejournal.org/issues/?fa=article&abstrID=7300>. Access in: Dec. 2019.

SUSAC, M.; SARLIJA, N.; HAS, A.; BILANDZIC, A. Predicting company growth using logistic regression and neural networks. *Croatian operational research review*. Croatia, v. 7, p. 229-248, 2016. DOI: https://doi.org/10.17535/crorr.2016.0016.

SUSEP - SUPERINTENDÊNCIA DE SEGUROS PRIVADOS. *Circular SUSEP Nº 145 de Novembro de 2.000. Dispõe sobre a estruturação mínima das Condições Contratuais e das Notas* Técnicas Atuariais dos Contratos exclusivamente de Seguros de Automóvel [. . .]. Rio de Janeiro: SUSEP, 2000. Available in: <http://www2.susep.gov.br/bibliotecaweb/docOriginal.aspx?tipo=1&codigo=9058>. Access in: Jan. 2020.

SUSEP - SUPERINTENDÊNCIA DE SEGUROS PRIVADOS *AuToseg*: sistema de estatísticas de automóveis da Susep. 2020. Available in: <http://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx>. Access in: Oct. 2019.

TANTITHAMTHAVORN, C.; HASSAN, A.; MATSUMOTO, K. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Access*, Guangzhou, 2018. Available in: <https://arxiv.org/pdf/1801.10269.pdf>. Access in: Apr. 2020.

YAO, L.; ZHONG, Y.; WU, J.; ZHANG, G.; CHEN, L.; GUAN, P.; HUANG, D.; LIU, L. Multivariable Logistic Regression And Back Propagation Artificial Neural Network To Predict Diabetic Retinopathy. *Diabetes Metab metabolic syndrome and obesity*, [Auckland], n. 12, p. 1943-1951. 2019. DOI: https://doi.org/10.2147/DMSO.S219842.

ZEMIACKI, J. *Teoria da credibilidade: Uma abordagem Bayesiana para estimação de prêmios de seguros de vida*. 2006. Trabalho de Conclusão de Curso (Bacharelado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006. Available in: <https://lume.ufrgs.br/handle/10183/134331>. Access in: Fev. 2020.

ZHANG, J.; CHEN, L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Computer Assisted Surgery*, Abingdon, v.24, n. 2, p. 62-72. 2019. DOI: https://doi.org/10.1080/24699322.2019.1649074.

Received: Apr. 22, 2020
Accepted: June 4, 2020