DOI 10.20544/HORIZONS.A.26.3.20.P07
UDC 368.212:303.22(497.16)
368.032:303.22(497.16)

REGIONAL ANALYSIS OF CAR INSURANCE IN MONTENEGRO USING DEEP LEARNING METHODS

Vladimir Kašćelan
University of Montenegro, Faculty of Economics, Podgorica,
vladok@ucg.ac.me

Ljiljana Kašćelan
University of Montenegro, Faculty of Economics, Podgorica,
ljiljak@ucg.ac.me

Milijana Novović Burić
University of Montenegro, Faculty of Economics, Podgorica,
mnovovic@ucg.ac.me

ABSTRACT

This paper analyzes the regional distribution of the number of policies, total premiums and claims of 19486 legal entities as car insurance customers of one of the largest non-life insurance companies in Montenegro. Insureds are clustered using the k-means methods in four clusters: Poor-low risk, Middle-low risk, Wealthy-middle risk and Luxury-high risk. The clusters are described using the Decision Tree (DT) models. Then, an analysis of the distribution of these clusters by regions is made. The results show that in the southern region, compared to the other two, there are fewer insureds who have a small number of policies, low premiums and low claims (Poor-low risk cluster), while more are from the Middle-low risk cluster. Also, the cluster of insureds with high premiums and high claims is most prevalent in the northern region. It is shown that deep learning methods can be used efficiently for this kind of analyses.

KEY WORDS: car insurance, deep learning, k-means clustering, decision tree, Montenegro

INTRODUCTION

Insurance companies have a strong interest to use a large amount of customer information applying various analytical techniques. Thus, they receive useful information for risk prediction, fraud detection, cost reduction, developing effective business, including marketing strategies to gain an edge over competitors. Boodhun (2017) gives an overview of several papers from 2013 to 2016 year dealing with data analytics in the insurance sector, outlining the machine learning approaches and algorithms used. Clustering, classification, neural networks and regression are the most popular approaches and techniques in machine learning methods.

According to Du (2006), insurers typically view the customer as an isolated object and having value only when insured communicates with them, neglecting the network value of each customer. Data Mining (DM) clustering techniques develop descriptive models to identify similar characteristics of policyholders in the same group and different characteristics in relation to other groups. A different marketing plan is created for each segment, instead of having a unique plan or plan for each policyholder individually. MergaGutema (2016) segments the clients of an Ethiopian insurance company,identifying the valuable segments of customers underlying variations of the customers' values. He points out that effective Customer Relationship Management (CRM) in companies should have a mechanism to increase the value of their customers. But some insurers are unable to succeed in this domain due to customers' related problems of satisfaction, retention and acquisition. In order to retain and maximize the profitability of their existing customers, it is important that these problems be resolved. Golmah (2014) deals with customer segmentation in the Iranian automobile insurance industry considering 23 companies and customer companies' choices in order to better understand insureds and increase customer loyalty and enhance customer retention. Roodpishi and Nashtaci (2015) extract hidden useful knowledge patterns and rules in the data set of one Iranian car insurance company and predict customer behavior.Zhuang et al.(2018)analyse characteristics of the customers of a car insurance company since 2014. It is used a business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms and a total of 25738 objects. Each of them is described with 46 attributes out of which twelve valuable attributes are chosen for clustering analysis.

In this paper DM techniques are applied in order to build segments made up of 19486 data and to categorize cascoinsureds of one Montenegrin insurer based on customer value. It will be shown that these homogeneous segments by region can be successfully identified using deep learning methods. The analysis of these segments provides significant information on the behavior of policyholders with similar characteristics in individual regions.

MATERIALS AND METHODS

This paper uses data from the database of one of the leading non-life insurance companies in Montenegro. Casco insureds, 19486 legal entities, are taken into consideration.

Table 1 shows the distribution of starting data for the analysis.

Table 1. Variables distribution and statistics

Variable	Statistics	Range
POLICIES	avg = 2.790 +/- 1.623	[1.000 ; 70.000]
PREMIUMS	avg = 803.219 +/- 1374.400	[10.900 ; 122577.170]
CLAIMS	avg = 331.567 +/- 1527.312	[0.000 ; 58520.940]
REGIONS	mode = Central (12787), least = North (921)	Central (12787), South (5778), North (921)

Table 2 shows the number of policies, total premiums and claims by regions, as well as the ratio of claims to premiums.

Table 2. Regional distribution of policies, premiums and claims

REGION	POLICIES	PREMIUMS	CLAIMS	CLAIMS vs.PREMIUMS
Central	34,118	10,083,021.58 €	4,354,938.90 €	43.19%
North	2,434	843,667.81 €	309,177.63 €	36.65%
South	17,813	4,724,827.61 €	1,796,800.44 €	38.03%
Grand Total	54,365	15,651,517.00 €	6,460,916.97 €	41.28%

Figure 1 shows the percentage distribution of policies, premiums and claims by regions.

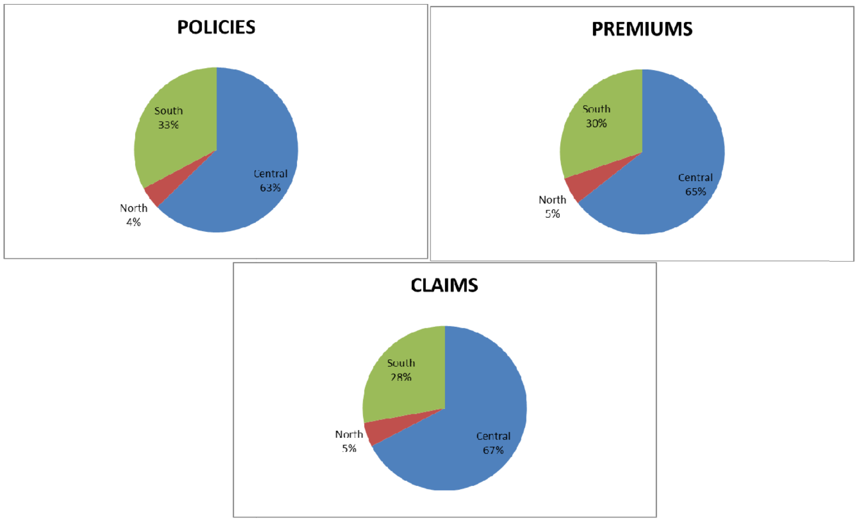

Figure 1. Regional representation of the policies, premiums and claims in percentages

It is important for insurance companies to determine the profiles of customer segments based on their behavior as policyholders and the distribution of these clusters across regions in order to create future business strategies. For this reason, the main objective of this paper is to show that deep learning methods can identify and describe segments of policyholders with similar behavior.

The k-means clustering method is applied to divide the insureds into homogeneous groups. This method is based on the geometric interpretation of data with n attributes as points in the n-dimensional space and the Euclidean distance between these points. In the initial iteration arbitrary k points are declare centroids and then the distance of one point to each centroid is calculated. This point joins to the centroidwith the least distance. In this way, each point is joined to one of the k centroids. Each centroid has the set of closest points to him that forms its cluster. In the next iteration, the mean of the coordinates of all points belonging to the cluster is calculated and the point thus obtained forms the new centroid. The procedure is repeated until maximum homogeneity within the cluster and maximum distance between clusters is achieved, what is being measured by an indicator known as the Davies-Bouldin(DB) index (MacQueen, 1967). The smaller the DB index, the better the clustering performance.

In order to describe the obtained clusters the classification DT method is applied. This method splits the starting dataset by the values of all attributes one after the other. The best division is taken, i.e. the one which gives the purest subsets with respect to the target classes (there should be as many representatives of one class as possible in each of the subsets obtained). Whether or not a division by someone attribute is good is determined by measures for entropy (information gain and gain ratio) or purity (Gini index) of the obtained subsets (Breiman et al., 1984; Quinlan, 1993). The resulting subsets are further subdivided until subsets of satisfactory purity are obtained. During partitioning, a tree graph is induced, where the nodescorrespond to the attributes selected for division and the edges correspondto the values of these attributes. Each leaf in this graph comprises a subset of the data obtained by the final division and represents the class that prevails in that subset. Root-to-leaf paths represent the rules by which classification is done. The more representatives of the class represented by the leaf, the greater the accuracy of the classification rule corresponding to that class. The number of accurately classified examples divided by the total number of data represents the accuracy of the model. The number of accurately classified examples of one class relative to the total number of examples that the model classified into that class represents class precision. The number of correctly classified examples of a class divided by the total number of examples belonging to that class is class recall. All these indicators are essential for the validity of the resulting DT model and for the classification rules. In this paper, the Gini index is used as a measure of quality of division, the classes are represented by clusters and the splitting attributes are policies, premiums and claims.

RESULTS AND DISCUSSION

The first step of the analysis is to divide clients into homogeneous segments based on their characteristics as insureds, i.e. based on their total number of policies as well as their total premiums and claims. Using the k-means method to the normalized data (0-1 transformation), the DB index is tested for the parameter k withvalues in range from 2 to 10. The results are shown in Table 3.

Table 3. Clustering performance

k	2	3	4	5	6	7	8	9	10
DB index	0.623	0.641	0.595	0.648	0.902	0.870	0.860	0.841	0.852

The table above shows that the DB index has the smallest value for k=4, which means that the division into four clusters is the best. The cluster model corresponding to this division is given in Table 4.

Table 4. Cluster Model

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
POLICIES	0.012	0.053	0.031	0.031
PREMIUM	0.005	0.009	0.014	0.027
CLAIM	0.002	0.004	0.109	0.440

Analyzing the normalized centroid values in the table above, it can be concluded that the insureds belong to one of the four clusters shown in Table 5.

Table 5. Clusters of insureds

	POLICIES	PREMIUMS	CLAIMS	Name
cluster_0 (12438)	Low	Low	Low	Poor-low risk
cluster_1(6577)	High	Low to Middle	Low to Middle	Middle-low risk
cluster_2(443)	Middle	Middle	Middle	Wealthy-middle risk
cluster_3 (28)	Middle	High	High	Luxury-high risk

Note: The number of policyholders within the clusters is shown in parentheses

As can be seen from Table 5 Cluster_0, named Poor-low risk, consists of 12438 policyholders who pay a small premium, have a small number of policies and low claims. Analogously, the Middle-low risk cluster consists of

6577 insureds paying low to middle premiums, have a high number of policies and low to middle claims, etc.

The next objective is to describe the identified clusters in more detail using the DT model. The generated DT model, shown in Figure 2, has a 99.9% classification accuracy, mean class precision and mean class recall, which confirms its validity.

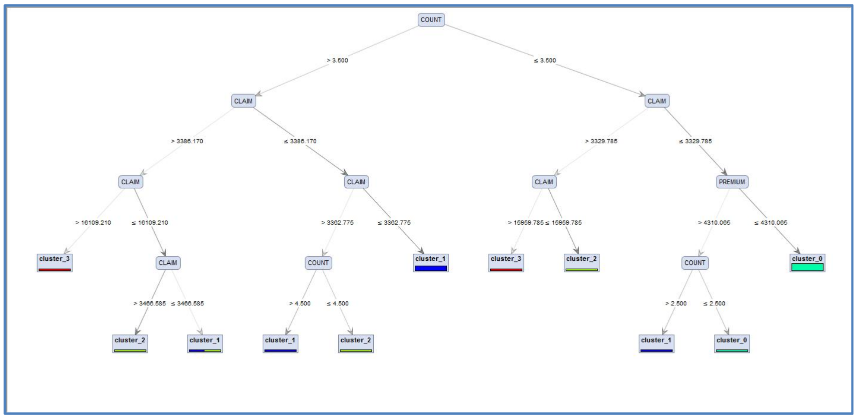

Figure 2. DT model for the description of the clusters of insureds

The classification rules derived from this model as well as the accuracy of each of the rules are shown in Table 6.

Table 6. If-Then rules for the clusters of insureds

Rule	Rule Accuracy
if COUNT > 3.500 and CLAIM > 3386.170 and CLAIM > 16109.210 then **cluster_3** (0 / 0 / 0 / 13)	100%
if COUNT > 3.500 and CLAIM > 3386.170 and CLAIM ≤ 16109.210 and CLAIM > 3466.585 then **cluster_2** (1 / 0 / 177 / 0)	99.44%
if COUNT > 3.500 and CLAIM > 3386.170 and CLAIM ≤ 16109.210 and CLAIM ≤ 3466.585 then **cluster_1** (4 / 0 / 4 / 0)	50%
if COUNT > 3.500 and CLAIM ≤ 3386.170 and CLAIM > 3362.775 and COUNT > 4.500 then **cluster_1** (3 / 0 / 0 / 0)	100%
if COUNT > 3.500 and CLAIM ≤ 3386.170 and CLAIM > 3362.775 and COUNT ≤ 4.500 then **cluster_2** (0 / 0 / 2 / 0)	100%
if COUNT > 3.500 and CLAIM ≤ 3386.170 and CLAIM ≤ 3362.775 then **cluster_1** (6526 / 0 / 0 / 0)	100%

if COUNT ≤ 3.500 and CLAIM > 3329.785 and CLAIM > 15959.785 then **cluster_3** (0 / 0 / 0 / 15)	100%
if COUNT ≤ 3.500 and CLAIM > 3329.785 and CLAIM ≤ 15959.785 then **cluster_2** (0 / 4 / 254 / 0)	98.45%
if COUNT ≤ 3.500 and CLAIM ≤ 3329.785 and PREMIUM > 4310.065 and COUNT > 2.500 then **cluster_1** (39 / 0 / 0 / 0)	100%
if COUNT ≤ 3.500 and CLAIM ≤ 3329.785 and PREMIUM > 4310.065 and COUNT ≤ 2.500 then **cluster_0** (1 / 35 / 0 / 0)	97.22%
if COUNT ≤ 3.500 and CLAIM ≤ 3329.785 and PREMIUM ≤ 4310.065 then **cluster_0** (3 / 12399 / 6 / 0)	99.93%

The DT model shows that the Luxury-high risk cluster consists of insureds whose number of policies exceeds 3 and whose claims are greater than 16109 euros, as well as those with number of policies less than 4 and whose claims are greater than 15959 euros. Insureds from Wealthy-middle risk cluster also have more than 3 policies butclaimsare between 3466€ and 16109€,own 4 policies and claims are between 3362 and 3386€, as well as those whose number of policies is less than 4, but claims are between 3329 and 15959 euros.

There are 4 rules associated with Middle-low risk cluster. The rule with the largest number of examples covered (6526 out of 6577)specifies that this cluster consists of insureds with more than 3 policies and claims less than 3362€ and with 100% rule accuracy. With significantly fewer examples covered by these rules, but with 100% accuracy, are the following: the number of policies is greater than 4, the claims between 3362 and 3386€, and the rulethat the number of policies is 3, each individual claim is less than 3329€ and a premium is greater than 4310€. The remaining rule has an accuracy of only 50% and the number of examples it covers is only 4.

Poor-low risk cluster is linked to 2 rules. More significant is the last rule in Table 6 (covers a large number of examples (12399) and has high accuracy (99.93%): these insureds have less than 4 policies, claims less than 3329€ and premiums less than 4310€.The DT model shows that the Poor-low risk cluster also consists of insureds whose number of policies is 1 or 2 and whose claims are less than 3329€, but premiums are greater than 4310€.

In addition to insureds segmentation, the aim of this research is to determine the regional distribution of individual segments, which can be of benefit to insurers for risk analysis and business strategy creation. Table 7 shows the regional distribution of the number of insureds by clusters and in Figure 3 shows the percentage representation of clusters for all regions and also by regions.

Table 7. Regional distribution of insureds by clusters

REGION	cluster_0	cluster_1	cluster_2	cluster_3	Grand Total
Central	8575	3903	295	14	12787
North	615	284	19	3	921
South	3248	2390	129	11	5778
Grand Total	12438	6577	443	28	19486

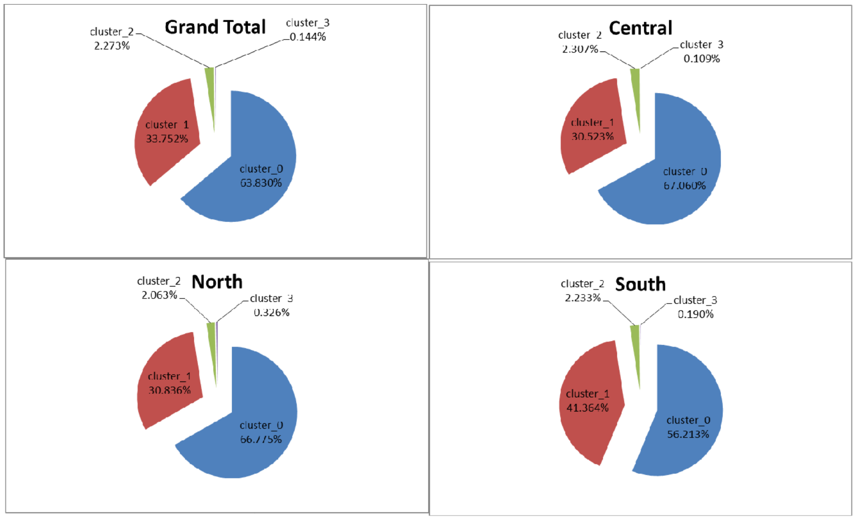

Figure 3. Percentage representation for clusters of insureds by regions

The Figure 3 shows that 63.83% of the total number of respondents is from the Poor-low risk cluster. Its share in the Central region is 67.06%, in North 66.775% and the lowest in South- 56.213%. The Middle-low risk cluster share is the largest in the South region- 41.364% and for the Luxury-high risk cluster in the North region- 0.326%. The share of the Wealthy-middle risk cluster (middle number of policies, middle claims and premiums) is uniform and ranges from 2.06% to 2.3%. In the southern region, compared to the other two, there are fewer insureds having a small number of policies, low premiums and low claims (Poor-low risk cluster), while more are from the Middle-low risk cluster.

Figure 4 shows the distribution of regions by clusters, where it can be seen that, observed isolated, the northern region isin the highest percentage (10.71%) involved in the cluster consist of luxury and more risky insureds (with high claims), than in other clusters, where its share is uniform and ranges from 4.29% to 4.94%.

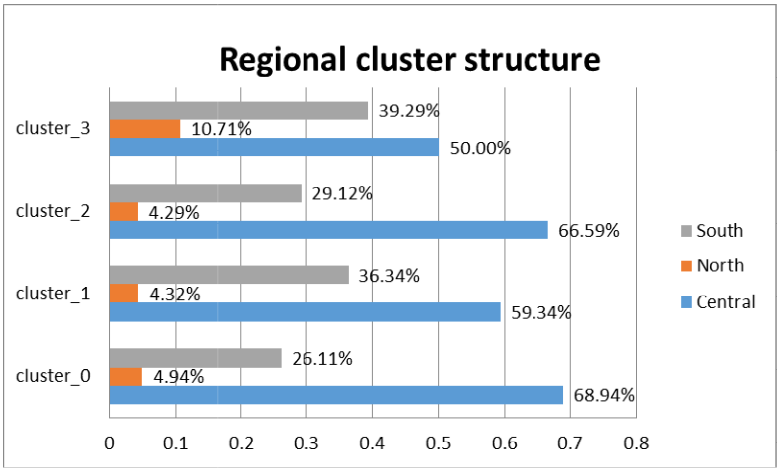

Figure 4. Percentage distribution of regions in clusters of insureds

The results show that homogeneous segments of the insureds can be efficiently identified using k-means and DT methods. Analyzing the distribution of these segments in different regions provides significant information about insureds and their behavior, which enables insurance companies to better assess risk and to communicate more efficiently with these market segments. Therefore, it can be concluded that the basic aims of this research are realized.

CONCLUSION

The aim of the study was to show that homogeneous segments of cascoinsureds, across the three regions of Montenegro, can be successfully identified using deep learning methods. It has been shown that clustering and classification approaches, i.e. the k-means and DT methods can be successfully applied to this insurance data in order to identify the profile of the insured.

Namely, for Poor-low risk cluster the most significant rule (covering a large number of examples and having high accuracy) shows that it is characterized by insureds with less than 4 policies, claims less than 3329€ and premiums less than 4310€. Based on the most significant rule regarding Middle-low risk cluster it is consists of insureds with more than 3 policies and claims less than 3362€ with 100% rule accuracy.The northern region is in the highest percentage involved in the Luxury-high riskcluster. The results show that in the southern region, compared to the other two, there are fewer insureds belonging to the Poor-low risk cluster, while more are from the Middle-lowrisk cluster.

The research results are of particular importance to the insurance company, i.e. its marketing and actuary departments, in order to establish better communication, to develop more successful marketing strategies and to better assess the risk by region for individual groups of policyholders of similar characteristics.

REFERENCES

1. Boodhun, N. (2017). A Review of Data Analytical Approaches in the Insurance Industry. *Journal of Applied Technology and Innovation*, *1*(1), 58-73.
2. Breiman, L., Friedman, J.C., Stone, J., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press
3. Du, X., (2006). Data Mining And Modeling For Marketing Based Attributes Of Customer Relationship. *Vaxjo University. Sweden Msi Report*
4. Golmah, V. (2014) A Case Study of Applying SOM in Market Segmentation of Automobile
5. Insurance Customers. *International Journal of Database Theory and Application.7(1). p.25-36.*
6. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1*, p.281–297.
7. Merga Gutema, D. (2016). *Application Of Data Mining Techniques For Customer Segmentation In Isurance Business: The Case Of Ethiopian Insurance Corporation* (Doctoral dissertation, Addis Ababa University)
8. Roodpishi, M., & Nashtaei, R. (2015). Market basket analysis in insurance industry. *Management Science Letters*, *5*(4), 393-400.

9. Quinlan J.R., (1993). *C4.5: programs for machine learning" vol. 1*. Morgan Kaufmann
10. Zhuang, K., Wu, S., & Gao, X. (2018). Auto Insurance Business Analytics Approach for Customer Segmentation Using Multiple Mixed-Type Data Clustering Algorithms. *Tehnički vjesnik*, *25*(6), 1783-1791.