INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 1(47) • 2018
ISSN 1507-3858
e-ISSN 2450-0003

Julian Vasilev, Miglena Stoyanova, Emiliya Stancheva
University of Economics – Varna, Varna, Bulgaria
e-mails: vasilev@ue-varna.bg; m_stoyanova@ue-varna.bg; emiliya.stancheva@ue-varna.bg

APPLICATION OF BUSINESS INTELLIGENCE METHODS FOR ANALYZING A LOAN DATASET

ZASTOSOWANIE METOD *BUSINESS INTELLIGENCE* DO ANALIZY DANYCH POŻYCZKOWYCH

DOI: 10.15611/ie.2018.1.08
JEL Classification: C45

Summary: The application of business intelligence methods is usually oriented on a specific type of business. This article is focused on the business intelligence data analysis of a sample loan dataset. There are a lot of well-known methods for analysing loan datasets. The research aim is to find some dependencies in a sample loan dataset which are not visible using techniques like sorting, filtering and pivoting. The following software products are used: Alyuda Neurointelligence and MS Power BI. As a result of the analysis it is proved that the amount of a loan depends mainly on the "born in the region" and "gender" factors. The analysis of the loan dataset in Alyuda Neurointelligence (using Input Importance) shows that the factor "born in the region" is the most important one affecting the amount of a loan. But the input importance shows only the possible strength of the influence of various factors on the amount of the loan and not the direction of their influence. The direction of their influence is found by applying the one-way ANOVA method in PSPP. The analysis of the loan dataset in Power BI shows some interesting dependencies. For debtors born outside the region, the difference in total loan amounts rises to over 4.5 times in favour of men. The average loan amount of debtors who were born outside the region is about 2.5 times greater than the average loan amount of debtors born in the region.

Keywords: Alyuda Neurointelligence, MS Power BI, PSPP.

Streszczenie: Dobór i wykorzystanie metod *business intelligence* do analiz danych jest zwykle zależne od typu prowadzonej działalności gospodarczej. Niniejszy artykuł przedstawia przykładowe zastosowanie metod *business intelligence* w analizach danych pożyczkowych. Istnieje wiele dobrze znanych metod analizy danych. Celem prowadzonych badań jest znalezienie pewnych zależności w przykładowej próbce danych pożyczkowych, które to nie są możliwe do przedstawienia za pomocą technik sortowania, filtrowania i obracania. Do przeprowadzenia analiz wykorzystano Alyuda Neuroinelligene oraz MS Power BI. W wyniku przeprowadzonych analiz udowodniono, że kwota pożyczki zależy głównie od czynników „urodzony w regionie" oraz „płeć". Wyniki analizy w Alyuda Neurointelligence (przy zastosowaniu ważności danych wejściowych, *Input Importance*) wykazało, że czynnik „urodzony w regionie" jest najważniejszy w definiowaniu poziomu wysokości udzielanej pożyczki. Jednakże ważność danych wejściowych wykazała tylko możliwą siłę wpływu

różnych czynników na wysokości pożyczki, a nie kierunek wpływu. Kierunek wpływu ważności danych wejściowych można znaleźć, stosując metodę jednokierunkową ANOVA w PSPP. Analizy danych pożyczkowych w MS Power BI wykazały wiele ciekawych zależności. Dla dłużników urodzonych poza regionem różnica w całkowitej kwocie posiadanych pożyczek wzrasta ponad 4,5 razy na korzyść mężczyzn. Średnia kwota udzielanych pożyczek dla dłużników urodzonych poza regionem jest około 2,5 razy większa niż średnia kwota pożyczek dłużników urodzonych w regionie.

Słowa kluczowe: Alyuda Neurointelligence, MS Power BI, PSPP.

1. Introduction

Dependencies may be found by statistical methods (such as correlation analysis) or by artificial intelligence methods such as a neural network with a multilayer perceptron. There are many techniques in statistics [Pallant 2011] which may be used to explore dependencies among variables.

Each credit institution has its own approach in estimating debtors. Risk analysis is a common technique. The result of risk analysis concerns some parameters of a loan. One of the most important parameters is the amount of the loan, when the potential debtor applies for a loan. The required amount depends certainly on the incomes of the debtor, but we are convinced that other factors affect the amount of the loan. That is why we try to find hidden dependencies in a sample dataset. The amount of the loan usually depends on the purpose of the loan. But in some cases the age of the debtor, the credit history of the debtor and his/her incomes and outgoings may also affect the amount loaned.

The main research problem is finding dependencies in a dataset using several software products (statistical software and BI software). No initial assumptions about dependencies in the dataset are defined. Using BI software we try to find some dependencies in the dataset – the possible factors affecting the amount of a loan. The possible dependencies (generated by BI software) are checked by an expert – if they are ambiguous or significant, after which statistical methods are used to find the mean difference in the loan amount among certain groups of people.

2. Literature review

Credit risk rating evaluation is specific for each country, for each economy sector and each credit institution. Some countries adapt national credit risk rating systems. A group of authors [Hai et al. 2013] propose a method for distinguishing default from non-default customers applying for farmer loans. The classification of customers is based on an empirical study using several key indicators: the purpose of the loan, the expenses/incomes ratio and regional GDP. Empirical studies about the loan lending policy of banks may be found [Delia, Luminita 2012]. These authors state that credit risk analysis is done mainly by using information about cash flows. They

prove that even though a strong mathematical formalization of credit risk analysis exists, when the economic environment is uncertain it is difficult to perform a reliable credit risk analysis.

Loan payment models are widely discussed. Two authors [Aydemir, Erdal Ramazan 2014] try to create a loan payment model with a linear gradient series. Several numeric examples are given to illustrate the proposed new loan payment model. Loans usually start with fixed payments, but later on some customers want to change some of the parameters of the loan. When they have payment months and non-payment months, new models are needed to describe the new parameters of the loan.

In statistics, causal dependencies are studied [Roberto et al. 2014]. Causal dependencies are usually obvious or they are proved in previous research articles. When creating econometric models several independent variables are defined and one or several dependent variables are also defined. Usually causal dependencies between variables are assumed, but using artificial intelligence methods [Vasilev, Atanasova 2015; Vasilev et al. 2017; Kuyumdzhiev 2016] new dependencies may be found. These dependencies may not have a visible causal relationship in this case statisticians usually call these dependencies "strange". Strange dependencies are usually found using data mining techniques.

3. Material and method

The dataset for the analysis is part of a sample dataset. Extract, transform and load (ETL) procedures are carried out to retrieve some of the columns from the whole dataset. All records (N = 712) are used. Some simple procedures for extracting the gender and place of birth from the civic number are done. The main assumption of this paper is that using business intelligence methods, some hidden dependencies may be found. This means that they are not clear using traditional methods for sorting, filtering and pivoting datasets. The columns of the dataset are the following: (1) month of signing the contract, (2) age, (3) gender, (4) born in the region and (5) amount.

The dataset is analyzed with Alyuda NeuroIntelligence (Alyuda Research n.d.) and MS Power BI.

4. Results and discussion

4.1. Analyzing the dataset in Alyuda NeuroIntelligence

The dataset has been imported into Alyuda NeuroIntellegence and 5 columns and 712 rows have been analyzed. From them, 5 columns and 698 rows have been accepted for neural network training and 14 rows are disabled.

The category columns are "gender", "month of signing the contract" and "born in the region". The numeric columns are "age" and "amount of the loan". The output column is "amount".

The aim is to find dependencies between the dependent variable "amount" and the other variables. A multilayer perceptron is used. A search for the most suitable architecture is conducted. The architecture [4-8-1] is chosen since it has the highest value of the fitness function – 0.016134 (Figure 1).

Architecture Search

ID	Architecture	# of Weights	Fitness	Inverse Test err	Akaike's criterion	R-Squared	Correlation	Train Error
1	[4-1-1]	7	0.013152	76.031265	-0.001184	0.066952	0.258752	78.435356
2	[4-10-1]	61	0.015585	64.163834	-0.001255	0.212049	0.463993	69.112885
3	[4-6-1]	37	0.016019	62.423981	-0.001134	0.259044	0.521792	63.889851
4	[4-4-1]	25	0.014337	69.748062	-0.00121	0.096877	0.318055	75.460281
5	**[4-8-1]**	**49**	**0.016134**	**61.979427**	**-0.001178**	**0.210118**	**0.476201**	**65.148735**
6	[4-9-1]	55	0.014777	67.672699	-0.001261	0.181283	0.427429	71.38475
7	[4-7-1]	43	0.014824	67.458618	-0.001223	0.163707	0.409344	71.291893

Fig. 1. The result of architecture search

Source: own study.

The neural network is created with Alyuda NeuroIntelligence using the Quick Propagation method. It has achieved a training error of 0.020118 and shows the input importance (Figure 2).

Network Statistics

Tabular Real-time

Input column name	Importance, %
month_contract	1.414124
age	20.772256
gender_1_female	9.12822
born_in_the_region	68.6854

Fig. 2. Level of input impact on the output using training with he quick propagation method (%)

Source: own study.

The result from statistics shows the following dependencies: the factor "born in the region" has the strongest impact on the amount of the loan (68.685%), after it the influence of other input variables is the following: age of debtors (20.772%) and gender (9.128%). The analysis shows that the month of signing the contract almost does not influence the loan amount (1.414%).

Using the Conjugate Gradient Descent method the impact of the examined inputs on the loan is the following: "born in the region" has again the strongest impact – 51.496%, the age – 25.835%, the gender – 14.095%, and last month contract – 8.572% (Figure 3).

Network Statistics

Tabular Real-time

Input column name	Importance, %
month_contract	8.572298
age	25.835017
gender_1_female	14.095743
born_in_the_region	51.496941

Fig. 3. Level of input importance on the output using training with the Conjugate Gradient Descent method (%)

Source: own study.

Using two other methods – Quasi-Newton and Limited Memory Quasi-Newton, we get approximately equal results – "born in the region" is the most important factor affecting the amount of a loan (95.252%, 96.834%), the gender has little or no effect (4.513%, 2.946%) and age and month of signing the contract have virtually no impact (Figures 4 and 5).

Network Statistics

Tabular Real-time

Input column name	Importance, %
month_contract	0.139513
age	0.094971
gender_1_female	4.513358
born_in_the_region	95.252158

Fig. 4. Level of input importance on the output using training with the Quasi-Newton method (%)

Source: own study.

Network Statistics

Tabular Real-time

Input column name	Importance, %
month_contract	0.114241
age	0.104117
gender_1_female	2.946704
born_in_the_region	96.834939

Fig. 5. Level of input importance on the output using training with the Limited Memory Quasi-Newton method (%)

Source: own study.

The created neural networks categorically proved that the factor "born in the region" is the most important factor affecting the amount of a loan. But the input importance shows only the possible influence of various factors on the amount of the loan, and not the direction of their influence. Therefore the results of an analysis of variance in PSPP (One Way ANOVA) and the values obtained by the neural network in Alyuda NeuroIntellegence can be compared.

A one-way ANOVA is conducted to compare the amounts for people born in the region and people not born in the region. There is a significant difference in the average amounts for people born in the region (N = 522, Mean = 79.67, Standard Deviation = 108.21) and people not born in the region (N = 190; Mean = 203.52, Standard Deviation = 346.02): F (1, 710) = 52.88, $p < 0.05$. The independent samples t-test shows similar results.

Training the neural network with different methods gives similar results. The results, obtained from the Conjugate Gradient Descent method, are the following: "born in the region" – 51.496%, "age" – 25.835%, "gender" – 14.095%, "month_ contract" – 8.572%.

4.2. Analysing the dataset in Power BI with quick insights and natural language processing

MS Power BI (https://powerbi.microsoft.com/en-us/) is a business intelligence tool which provides data analysis and interactive visualizations. The dataset is analyzed with Power BI using natural language processing (NLP). The outcomes show some interesting dependencies.

One of the questions that are asked in Power BI is "What is the average amount by gender?". The obtained values for the average loan amount by gender are 119.64 for men and 91.81 for women. We may expand the first question by adding the attribute "month". Thus the question becomes "What is the average amount by gender and by month?". The answer to this question is shown in Table 1 as a matrix. The average loan amount by gender and month may give information for the months with the highest average loan amounts for men and women. Significant differences in the values between the genders can be observed in March, June, August and October.

Figure 6 presents a visualization, obtained in response to the question "What is the average loan amount by month?" in MS Power BI. It can be seen that contracts signed in November have the highest average loan amount. The smallest average loan amount is in March.

The analysed data can be summarized by the column "born in the region" with the question "What is the total amount by born in the region?". The total loan amount for this indicator shows higher values for debtors, who were born in the region.

We can compare the total loan amounts for both genders, combined with the attribute "born in the region". For this purpose we asked MS Power BI the following question: "What is the total amount by gender and born in the region?" From the

Table 1. Average loan amount by gender and by month of contract

Gender	Month of contract	Average of amount
Female	August	45.22
	December	97.27
	February	107.83
	January	160.26
	July	95.37
	June	72.50
	March	44.44
	November	116.00
	October	60.83
	September	78.13
	Total	**91.81**
Male	August	89.00
	December	129.06
	February	105.97
	January	108.19
	July	95.49
	June	136.74
	March	77.47
	November	164.57
	October	146.15
	September	116.27
	Total	119.64
Total		112.72

Source: own study.

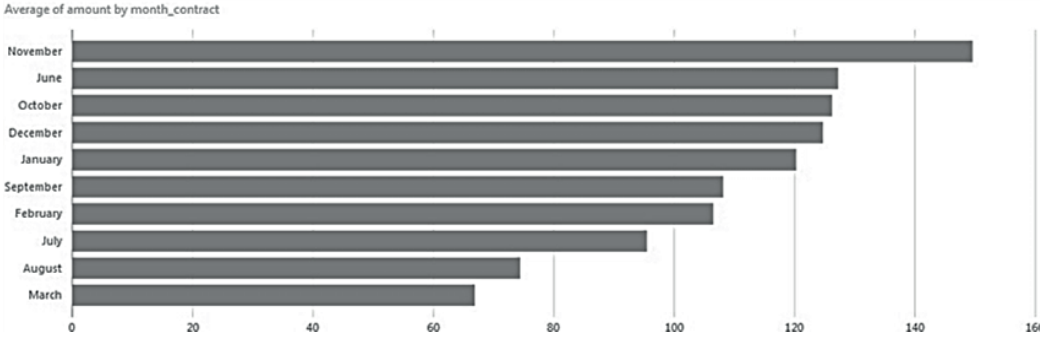

Fig. 6. Average amount by month

Source: own study.

results for all debtors who were born in the region of the credit institution, we can conclude that the men's total loan amount (32 228) is about 3.5 times greater than the women's total loan amount (9360). For debtors born outside the region, the difference in total loan amounts rises to over 4.5 times in favour of men (men – 31 777, women – 6891). Moreover, the total loan amounts for men born in the region and those born outside the region, have very similar values. At the same time, the total loan amount for the women is greater for those who were born in the region.

The following question is similar to the previous: "What is the average amount by gender and born in the region?". The analysis of the average loan amount by gender and debtor's region of birth shows that men (217.65) and women (156.61), who were born outside the region of the credit institution have higher average loan amounts. The corresponding average loan amounts for men and women who were born in the region, are 82.85 and 70.38. We may conclude that the average loan amount of debtors who were born outside the region is about 2.5 times greater than the average loan amount of debtors born in the region (see Figure 7).

Average of amount	born_in_the_region
203.52	no
79.67	yes
112.72	

Fig. 7. Average amount by "born in the region"

Source: own study.

Another interesting metric for comparison is the debtor's age. The three highest average loan amounts (600, 349.50 and 315.17) are for the ages of 53, 46 and 38. If we consider the average loan amounts by debtors' age and gender, we can conclude that the highest amounts are for men aged 53 and 38, and for women 46 and 41 years. These values may be interpreted as outliers, not as meaningful dependencies. All the received answers from MS Power BI in the form of charts and tables show outliers which are not found by using the Quick Insights tool in MS Power BI.

The analysis of the loan dataset in Power BI shows interesting information. Different factors which may affect the amount of the loan are presented. It is important to note that the obtained results should be validated by an expert, who may decide which dependencies can be useful for loan amount prediction and debtor's estimation.

5. Conclusion

Predicting the amount of a loan by finding dependencies in a sample dataset is a difficult task. The experts may use different software tools and business intelligence methods to support their work.

Several neural networks are tested in Alyuda Neurointelligence. Different training methods are tested. The analysis of the loan dataset in Alyuda Neurointelligence (using Input Importance) shows that the factor "born in the region" is the most important one affecting the amount of a loan. But the input importance shows only the possible strength of the influence of various factors on the amount of the loan and not the direction of their influence. The direction of their influence is found by applying the one-way ANOVA method in PSPP.

The analysis of the loan dataset in Power BI shows some interesting dependencies between the variables in the dataset. Contracts signed in November have the highest average loan amount. From the results for all debtors who were born in the region of the credit institution, we can conclude that the men's total loan amount is about 3.5 times greater than the women's total loan amount. For debtors born outside the region, the difference in total loan amounts rises to over 4.5 times in favour of men. The average loan amount of debtors who were born outside the region is about 2.5 times greater than the average loan amount of debtors born in the region. Different factors which may affect the amount of the loan and the debtor's estimation are found. The obtained results should be validated by experts who may decide which dependencies can be useful and meaningful in making a particular decision.

Future research may focus on finding other dependencies in loan datasets, and may focus on testing other models which better describe the dependencies between other covariates and the amount of a loan.

Bibliography

Alyuda Research, Alyuda_Neurointelligence. Available at: http://www.alyuda.com/neural-networks-software.htm [accessed January 20, 2016].

Aydemir, Erdal Ramazan E., 2014, Development of New Loan Payment Models with Piecewise Geometric Gradient Series, Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi., 29(2), pp. 95-106.

Delia D., Luminita P., 2012, *Methods used by banks when taking medium term and long term financing decisions*, Annale : Seria Stiinte Economice. Timisoara, 18, pp. 432-440.

Hai L., Shi B., Peng G., 2013, *A credit risk evaluation index system establishment of petty loans for farmers based on correlation analysis and significant discriminant*, Journal of Software, 8(9), pp. 2344-2351, available at: http://ojs.academypublisher.com/index.php/jsw/article/view/10139.

Kuyumdzhiev I., 2016, *The DIMBI project – innovative approaches for teaching business informatics*, Scientific Journal Economics and Computer Science, 2016, vol. 2, issue 5, pp. 26- 37, available: http://eknigibg.net/Volume2/Issue5/spisanie-br5-2016_pp.26-36.pdf

Pallant J., 2011, *SPSS SURVIVAL MANUAL: A step by step guide to data analysis using SPSS*, Allen and Unwin.

Roberto J., Cruz P., Hernandez S.E.P., 2014, *Temporal alignment model for data streams in wireless sensor networks based on causal dependencies*, International Journal of Distributed Sensor Networks, pp. 8-10.

Vasilev J., Atanasova T., 2015, *Parallel testing of hypotheses with statistical and artificial intelligence methods : A study on measuring the complacency from education*, Computer Science and Applications, 2(5), pp. 206-211.

Vasilev J. et al., 2017, *Business intelligence. Varna: Knowledge and business*, available: https://activetextbook.com/active_textbooks/13534.