|  |  |
| --- | --- |
|  | **Higher School of Economics**  **National Research University** |

**Case Study: Data mining and knowledge discovery**

Case prepared by Professor **Zinaida Avdeeva** and **Alexey Naydenov**

**Case description**

Eliot Ness - special agent of Ministry of Finance decided to improve the work of tax department in terms of counteracting tax evasion. Analytics team is appointed to develop methodology of investigating financial data of enterprises aimed to reveal those taxpayers who evade taxation.

Analytics team apply data mining which “is the core of the Knowledge Discovery process, involving the inferring of algorithms that explore the data, develp the model and discover previously unknown patterns. “[[1]](#footnote-1) In this case study you will follow the way of analytics team, which is in general includes two basic steps:

* obtain clusters of taxpayers according their economic sector, effectiveness of business and the scale;
* find outliers among clusters – those who do not fall in typical business structure of financial characteristics of enterprises.

The above mentioned way is based on the following research hypothesis – in the same economic conditions and having the same type of business financial indicators are similar and could be clustered in patterns. The outliers in the clusters are probably companies which have not typical economic activity: both legal and illegal. The first case is not the deal of tax agency but the second one is. Anyway, the point is to find those companies which have such not typical activities and investigate them.

**1. Import data**

You can import data from “data.csv” file.

**2. Data cleansing**

The main preliminary step of data mining is data cleansing which takes sometimes 80% of the whole process time, being important factor of success. You should explore the data for possible problems and endeavor to correct the errors.

Convert to factor variable string values. Please pay attention on the different spelling of the same parameter values due to imputation variation: “construction” = “Construc.”. Please study dataset for typos, they could bring to errors in further procedures.

Study summary statistics of data in order to exclude typos, NAs and outliers, caused by typos. For example, most of variables are obviously nonnegative. But be careful with data cleansing: some parameters could be really both positive and negative, while others are only positive or negative. Make assumptions on every parameter according to the summary statistics, boxplots, etc.

*The adequate outliers detections assumes accuracy in terms of the current research topic. Please distinct outliers in variables distribution (probably, a signal of tax evasion) and outliers as unreal values.*

a) find NAs and substitute them with 0;

b) find negative outliers, which are assumed to be caused by typos due to summary statistics or the nature of variable – capital is hardly to be negative in any cases of doing business.

c) illustrate the task and interesting findings by few examples of summary statistics, plots, charts, etc. Show your approach, describe a package or functions you have used.

Note: you should try a several approaches and packages: ***summary*** from base package, ***outliers*** or ***moments*** packages.

**3. Explore the data**

Check normality of data. You can use Shapiro-Wilk test (***shapiro.test*** function). Analyze kurtosis and skewness of data: we advise you to use ***moments*** package.

Give illustration of variables distributions: histograms and estimations of kurtosis and skewness.

The next step is investigating the relationship between two variables in terms of covariance and correlation (***cov*** and ***cor*** functions). A matrix of scatter plots can be produced with the function ***pairs***.

Further analysis could be done via ***scatterplot3d*** package which gives additional visualization tools and using ***heatmap*** function.

Investigate packages MASS, lattice, ggplot2 and others in terms of presenting the results of exploratory analysis. In context of current research multivariate analysis is not the topic of interest, but usually this part is very important.

**4. Clusterization**

Prepare the data for clustering. Pay attention on the real structure of enterprises, data contains enterprises from different economic sectors. Do clusters correlate with sectors?

There are two main approaches to clustering: K-means and hierarchical. Try both of them and compare results. Note: k-means algorithms assumes that you know the amount of clusters before clustering process. Try different amount of clusters (from 2 to 6) and determine the optimum. Compare the structure of clusters, have you get the same results?

Plot cluster solutions (packages ***cluster*** and ***fpc*** will help you).

Validating of cluster solutions is important part of clustering. The function ***cluster.stats*** in the ***fpc*** package provides a mechanism for comparing the similarity of two cluster solutions using a variety of validation criteria (Hubert's gamma coefficient, the Dunn index and the corrected rand index).

**5. Final. Find tax evaders**

The final steps includes revealing of tax evaders. According to the hypothesis of the research our task is to find outliers.

***a) three-sigma rule and interquartile ranges***

First of all you should check the normality of distribution of variables (Shapiro-Wilk test). If distribution is not normal, try to use standardization methods:

* min-max normalization;
* z-score.

You can use open sources to get formulas for calculation the normalized values. Does any of the method allows to transform data into normal distribution?

Usually, optimal solution gives BoxCox method, provided by several packages. We suggest to use ***forecast*** package, which contains function for picking up best lambda value (use ***BoxCox*** and ***BoxCox.lambda*** functions together).

After you have transformed the data you should detect the outliers using two different approaches: three-sigma rule and interquartile ranges.

The first one expresses a conventional heuristic that nearly all values are taken to lie within three standard deviations of the mean. And the second says that an outlier is any value that lies more than one and a half times the length of interquartile range below or above first and last quartile relatively.

Compare both approaches and illustrate the difference of revealed outliers.

***b) outlier detection via packages***

There are several packages to perform the task:

* outliers – tests for outliers;
* mvoutliers – multivariate outlier detection based on robust methods;
* extremevalues – univariate outlier detection.

We shall use Local Outlier Factor algorithm from packages ***Rlof*** and ***DMwR*** (packages ***dprep*** as an option). Rlof, comparing with the DMwR has two additional features, including several choices of distance metrics.

Please try both of them for outlier detection, compare the results between the packages and used distance metrics. Illustrate the results with plots and sets of outliers detected. Try to use both normalized and raw datasets. Have you gotten the same results?

**6. Report**

Please prepare a report and a presentation on the results of your research.

**References**

* Cotton, R. (2013). Learning R. O’Reilly. 2013.
* R and Data Mining: Examples and Case Studies. Y Zhao. Academic Press, 2012.
* Zuur, A. F., Ieno, E. N. and Elphick, C. S. (2010), A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1: 3–14. doi:10.1111/j.2041-210X.2009.00001.x
* Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10
* Edwin de Jonge, Mark van der Loo (2013). An introduction to data cleaning with R

1. Data Mining and Knowledge Discovery Handbook. Second Edition. Editors: Oded Maimon, Lior Rokach. Springer, 2010. DOI: 10.1007/978-0-387-09823-4 [↑](#footnote-ref-1)