Applying Data Mining Algorithms on Demographic Data with Power BI

Tsvetanka Georgieva-Trifonova*, Tihomir Trifonov**, Márta Seebauer***
* "St. Cyril and St. Methodius" University of Veliko Tarnovo, Department of Mathematics and Informatics, Veliko Tarnovo, Bulgaria
** Vasil Levski National Military University, Veliko Tarnovo, Bulgaria
*** Óbuda University, Budapest, Hungary
cv.georgieva@uni-vt.bg, TihomirTrifonov55@gmail.com, seebauer.marta@amk.uni-obuda.hu

***Abstract*— Analyzing business results along with demographic data is a useful approach to discovering important summaries, hidden dependencies, and predictions to support making successful business decisions. This paper examines the features of Power BI for performing clustering; correlation analysis; descriptive analysis by using decision tree. Illustrative examples about data, available on the website of the National Statistical Institute in Bulgaria are presented. The results confirm the benefits of applying these analyzes to demographic data.**

I. INTRODUCTION

Business companies store and continuously generate data from different domains related to their business: ads, sales; customers, users, feedback reviews; accounting data; text content, video data, photos, graphics, others. Data accumulation strategies may include tools for:

- analyzing customer needs and preferences to provide a better understanding and targeting the customers;
- managing and searching for consumer feedback, including from social networks;
- support ads and ad campaigns targeting, etc.

The gathered data provide huge opportunities for companies, but various researchers question whether the maximum benefit from them is extracted in fact. If the archiving data work strategy is presented as a human needs hierarchy defined by Maslow, the resulting pyramid is shown in Fig.1. According to [13], the level of media companies falls into the middle, between the stages of "Information" and "Knowledge".

Figure 1. Maslow's hierarchy presented the stages of working with big data [13]

In order to reach the "Wisdom" stage, companies need to take advantage of the work and data analysis capabilities. For this purpose, they have to invest in technology and training. The expected result is to improve processes such as those related to customer service and products.

The present research is motivated by the fact that the combination of demographic data with data from a particular business activity or customer feedback and the implementation of data mining algorithms allows to obtain aggregated information, forecasts, and dependencies, useful to support making adequate business decisions.

The advantages and disadvantages of implementing data mining algorithms through Power BI are summarized. Execution of clustering, correlation analysis, and descriptive analysis by tree decision on demographic data is demonstrated.

This paper is organized as follows. In Section 2, an overview of existing research to implement demographic data mining methods is made. Section 3 summarizes the advantages and disadvantages of applying data mining by using Power BI. Section 4 is dedicated to experimenting with demographic data downloaded from the National Statistical Institute's website in Bulgaria to derive the benefits that Power BI offers.

II. REVIEW OF EXISTING RESEARCH TO IMPLEMENT DEMOGRAPHIC DATA MINING METHODS

Clustering methods are successfully applied when analyzing demographic data [2, 6, 10, 11].

In [1], the methods of Holt and Holt-Winters for multiplier and additive seasonality are applied to obtain monthly forecasts of the average monthly salary (in BGN) of employees in Bulgarian tourism. The forecast is based on data from the National Statistical Institute on the average annual salary of employees under labor and employment relationships by economic activities and forms of ownership in the hotel and restaurant sector.

A correlation analysis is applied in [9] to explore the relationship between demographic data of Twitter readers (education, location) with their credibility perception. In order to improve the effectiveness of e-learning, a classification is applied to define a set of student profiles in [14]. For the classification, survey data are collected, some of which relate to demographic information, such as sex, age group, university, faculty, specialization, school year.

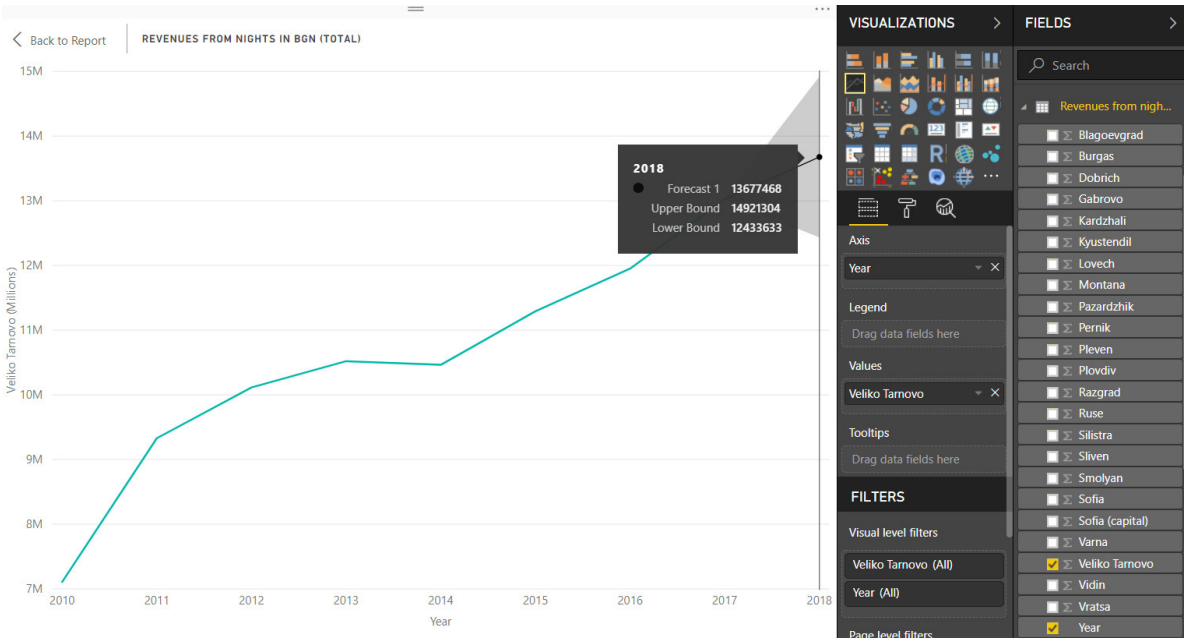


Figure 2. Overnight stays in Veliko Tarnovo District

Demographic sequences are examined and analyzed in [5, 8]. Comparing methods for demographic data classification is performed in [8]. Various methods such as decision trees, support vector machines, K-Nearest Neighbors (K-NN), neural networks are used for classification.

A product recommendation system is proposed in [12], which uses additional information, including demographic data about users, in order to improve the accuracy of the recommendations made.

In the present paper, the features of Power BI for implementing data analysis as clustering, correlation, analysis, descriptive analysis by tree decision are examined.

III. ADVANTAGES AND DISADVANTAGES OF APPLYING DATA MINING THROUGH POWER BI UNITS

Microsoft Power BI is a flexible tools for applying Business Intelligence, which facilitates users to self-generate data analytics managed in the cloud for the collaborative work and sharing [3]. This tool is distinguished from a data warehouse because it allows:

– *User-defined, flexible data retrieval and analysis;*

This includes creating calculated columns; changing measures and dimensions on-the-fly, and without the need for extra use of other IT and ETL (extracting, transforming, and loading) tools.

Typically, data warehouses require data to be imported (via ETL tools) into the warehouse;

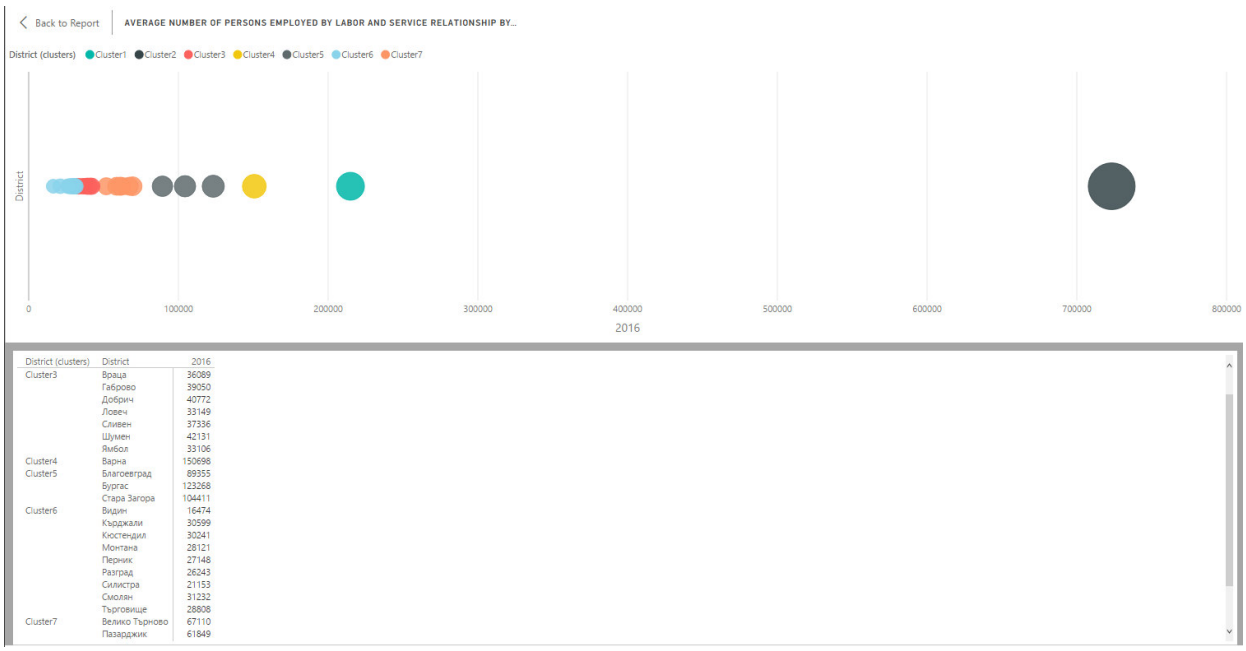
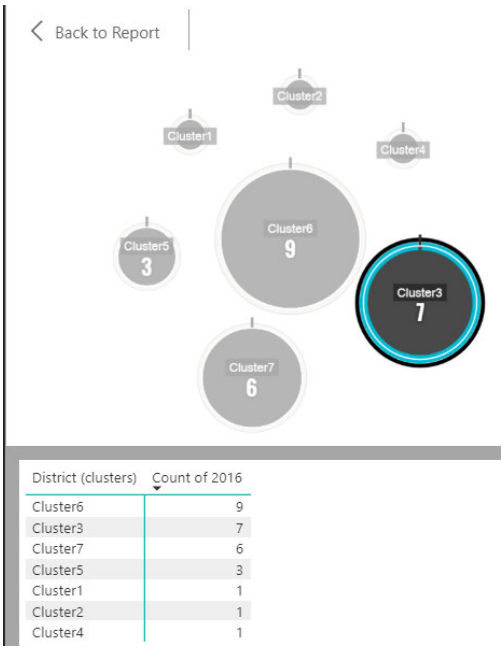


Figure 3. Clustering of the Regions According to the Number of Employers



District (clusters)	Count of 2016
Cluster6	9
Cluster3	7
Cluster7	6
Cluster5	3
Cluster1	1
Cluster2	1
Cluster4	1

Figure 4. The Clusters Visualized with Cluster Map

generating cubes with data; then analyze the data, thus making the BI process heavier, dependent on the IT technology used and difficult to repeat in short periods.

– *Working with a graphical interface;*

Online publishing in such a way that allowing data to be shared from a central location;

A. *Advantages of Power BI*

The benefits of Power BI can be summarized as follows:

(1) It provides the ability of analysts who can retrieve source data, create a dataset, transform or manipulate data, visualize data, and publish the resulting reports and dashboards;

(2) The basic Power BI technology is a database that supports table structures for data used by Power Pivot. By this way a compromise between performance and ease of use is achieved. On the other hand, data cubes require more complex structuring and query language, such as MDX (MultiDimensional eXpressions);

(3) The language DAX (Data Analysis Expressions) is based on a relatively simple statement used to create computed columns and measures;

(4) Q&A feature. This feature allows retrieving a response from the data by asking a question by performing natural language processing. Q&A is the most commonly mentioned benefit for self-service Business Intelligence;

(5) Actively maintained and updated. The available visuals are continuously updated by the community. Interactive visualizations of geographic maps are authorized by Bing Maps. Dynamics CRM integration also evolves;

(6) The free version reduces user limitations.

B. *Disadvantages of Power BI*

The following major limitations and disadvantages of Power BI can be identified:

(1) A given dataset may include multiple datatypes, but Power BI reports and dashboards can only generate data from one dataset; Similarly, Power BI can not combine imported data together with data available from real-time connections;

(2) Power BI can not work with files larger than 1 GB. The limitation refers to .pbix files, which are a type of archive files and compress data until they have to be used. However, the maximum allowed file size may limit Power BI to a subset of the data extracted from the data warehouse(s) of the organization concerned;

(3) For a dataset size, the limit is up to 1GB when using the free cloud service PowerBI.com. An example solution is to create multiple smaller datasets. For Pro version of PowerBI.com cloud service, the limit is up to 10GB;

(4) The granularity is not supported. Microsoft Power BI does not allow for creating scheduled reports, customized user views, notifications, and reports.

(5) In the presence of multiple different use cases and user profiles that require customization, Microsoft Power BI is not the best solution.

(6) Data cleaning. Microsoft Power BI has the ability to clean and transform imported data to achieve improvement of the data quality.

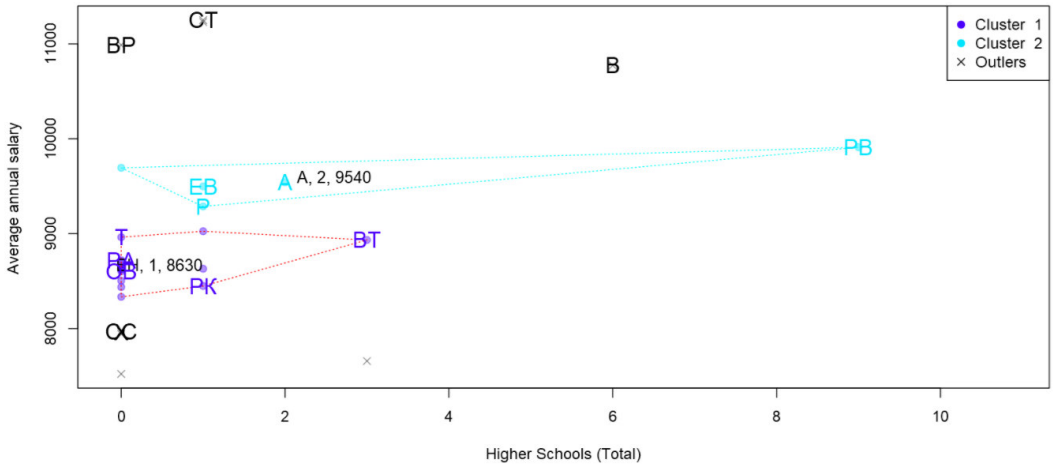


Figure 5. Clustering Based on the Average Annual Salary per District and the Number of Higher Education Institutions

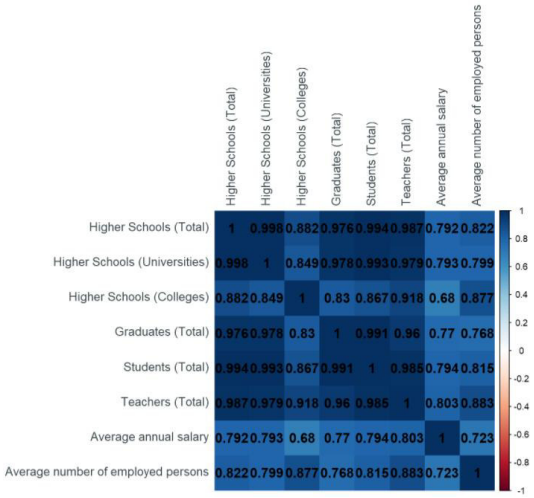


Figure 6. Complex Correlation Matrix

However, for more sophisticated transformations, it is necessary to use other tools such as MS SQL Server Integration Services or R script [4];

(7) This Microsoft tool is often used to supplement instead of replacing other reporting tools. In many cases, Power BI does not replace the corporate data warehouse. The tools for company data storage can be used for rarely changing reports when large-volume data processing is required. On the other hand, Power BI can be used for single and continuous, frequently changing analyzes of smaller datasets.

Based on the performed study of the main advantages and disadvantages of Power BI, we can conclude that the analysis of demographic data can be successfully done using this tool. Experimenting with specific data is described in the following section.

IV. Analyzing and Visualizing Demographic Data through Power BI

In addition to standard data representation in graphics and charts, Power BI allows the implementation of forecasting algorithms based on available archive data; hidden previously unknown dependencies in the data.

In this research, the analysis and visualization are applied on the demographic data obtained from the website of the National Statistical Institute in Bulgaria concerning

- the average number of employed persons;
- average salary;
- number of universities;
- number of university graduates by district, etc.

The visualizations used are:

- Cluster map;
- Correlation plot;
- Clustering with outliers;
- Decision Tree Chart.

A. *Forecasting*

Forecasting in Power BI is based on a method called exponential smoothing, which is a modification of the Holt-Winters method. The concrete example shown in Fig. 2 refers to overnight stays in the Veliko Tarnovo district at a 95% confidence interval.

B. *Clustering analysis*

Power BI Desktop provides the ability to automatically find clusters of data. Clustering allows groups of similar data to be discovered in the dataset.

Fig. 3 shows clustering of the regions according to the average number of persons employed by labor and service relationship by districts in 2016. As a result, seven clusters are obtained.

The visualization *Cluster Map* is designed to represent clusters of related data. An appropriate data summary can be selected – number of objects in each cluster (Fig. 4); sum; average; minimum, maximum value, etc.

The visualization *Clustering with Outliers* uses DBSCAN algorithm (*Density-Based Spatial Clustering of Applications with Noise*), which also finds the exceptions.

Fig. 5 depicts the clusters found, when the points being derived based on the average annual salary per district and the number of higher education institutions in the respective district.

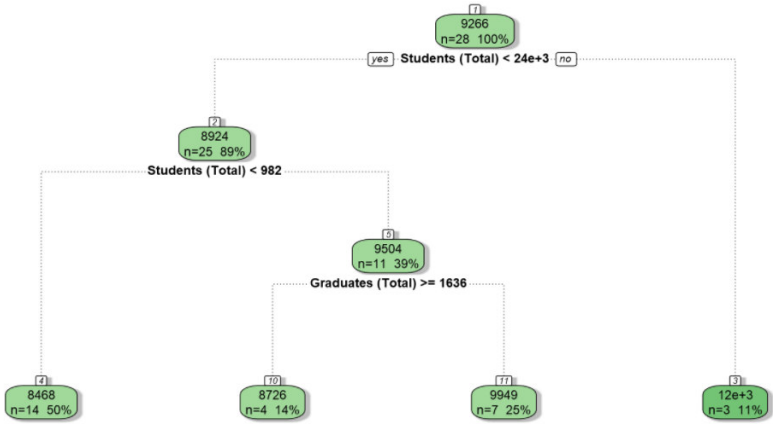


Figure 7. Correlation Tree between the Number of Students, the Total Number of Graduates and the Average Annual Salary per District

C. Correlation analysis

The visualization *Correlation Plot* is based on correlation analysis, a method that aims to find a relation between different attributes in a given dataset. The presence of a relation between two attributes means that when there is a change in one attribute, the other one is also changed. For this purpose, a correlation coefficient is calculated, i.e. the so-called Pearson coefficient [7], whose value is in the range –1 to +1.

The values close to:

–1 indicate a negative relationship, i.e. the high numerical values of an attribute refer to low numerical values of another attribute;

+1 indicate a positive relationship, i.e. the high numerical values of an attribute refer to high numerical values of another attribute;

0 – lack of correlation.

On Fig. 6, the correlation between the following attributes is examined: the number of higher schools; graduate number; number of students; number of lecturers; average annual salary; average number of persons employed by labor and service relationship by districts in 2016.

D. Decision tree

The visualization *Decision Tree Chart* implements a recursive splitting during the decision tree construction.

Fig. 7 shows the decision tree obtained by the following settings:

- the input variables are the total number of students and the total number of graduates in the respective fields;
- the target variable is the average annual salary per district.

Applying the presented visuals to demographic data combined with business data could provide descriptions and summaries of data; forecasts; dependencies that are helpful in supporting appropriate business decisions.

V. Conclusion

Advanced and easy-to-use tools are intensively developed and refined to perform complex data analysis. It's no accident that Gartner recognizes Microsoft as a leader in BI platforms for 11 consecutive years. The present paper explores the benefits of using Power BI to implement algorithms to mining demographic data and conveniently visualize the results obtained.

Acknowledgment

The work presented in this paper was partially supported by the research project “Analyses and research with cloud technologies for business”, “St. Cyril and St. Methodius” University of Veliko Tarnovo.

References

[1] Preslav Dimitrov, Can the Bulgarian Tourism solve the problem with poverty in Bulgaria?, In book: The Fight with Poverty in Bulgaria: Macroeconomics, Ethical Aspects, Eurointegration, the Role of TourismChapter, Fridrich Ebert Stiftung: Beureau Bulgaria & Institute for Economics and International Relations, Sofia, 2014.

[2] Nontembeko Dudeni-Tlhone, Jenny P. Holloway, Sibusisiwe Khuluse, Renée Koen, Clustering of Housing and Household Patterns Using 2011 Population Census, 55th Annual Conference of the South African Statistical Association, 2013.

[3] Esat Erkec, How to create advanced analytics using Power BI and R scripts, SQLShack, 2018, available at: https://www.sqlshack.com/use-advanced-analytics-power-bi/

[4] Oliver Frost, Power BI: Transform Data with R, 2017, available at: https://www.consolidata.co.uk/explore/blog/power-bi-transform-data-with-r/

[5] Dmitry I. Ignatov, Ekaterina Mitrofanova, Anna Muratova, Danil Gizdatullin, Pattern Mining and Machine Learning for Demographic Sequences, International Conference on Knowledge Engineering and the Semantic Web, 2015, pp 225-239.

[6] Elza Jurun, Nada Ratković, and Ivana Ujević, A cluster analysis of Croatian counties as the base for an active demographic policy, Croatian Operational Research Review, vol. 8, 2017, pp 221–236.

[7] Daniil Maslyuk, Pearson correlation coefficient in DAX, 2017, available at: https://xxlbi.com/blog/pearson-correlation-coefficient-in-dax/

[8] Anna Muratova, Pavel Sushko, Thomas H. Espy, Black-Box Classification Techniques for Demographic Sequences: from Customised SVM to RNN, In Proceedings of the Fourth Workshop on Experimental Economics and Machine Learning, Dresden, Germany, 2017, pp. 31-40.

[9] Shafiza Mohd Shariff, Mark Sanderson,and Xiuzhen Zhang, Correlation Analysis of Reader’s Demographics and Tweet Credibility Perception, 38th European Conference on IR Research, 2016, pp 453-465.

[10] Oksana Shubat, Abilova Makhabat, Anna Bagirova, Anton Ivlev, The Use Of Cluster Analysis For Demographic Policy Development: Evidence From Russia, Proceedings 30th European Conference on Modelling and Simulation, 2016.

[11] Ondřej Šimpach – Marie Pechrová, Cluster Analysis of EU’S Regions According to Demographic Criteria, The 10th International Days of Statistics and Economics, Prague, September 8-10, 2016.

[12] R. Satish Srinivas, C. S. Anish Balaji and P. Saravanan, Online Product Recommendation using Relationships and Demographic Data on Social Networks, Indian Journal of Science and Technology, Vol. 9, no. 44, 2016.

[13] M. L. Stone, Big Data for Media, Reuters institute for the study of journalism, University of Oxford, Report, 2014.

[14] Alexandru Topîrceanu, Gabriela Grosseck, Decision tree learning used for the classification of student archetypes in online courses, Procedia Computer Science, vol. 112, 2017, pp.51–60.