Advanced analytics of Big Data using Power BI: Credit Registry Use Case

Fisnik Doko
Faculty of Informatics
“Mother Teresa” University
Skopje, Macedonia
fisnik.doko@unt.edu.mk

Igor Miskovski
Faculty of Computer Science and Engineering
“SS. Cyril and Methodious” University - Skopje
Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

Abstract—**Big Data is emerging trend which brings and the need for more effective data analysis and visualization to get new knowledge and to leverage the benefits of advanced analytics of the volume of data they collect without IT knowledge. Analyzing and visualizing large volumes of data in financial services often suffers from performances in traditional systems with traditional tools. For understanding the data through visualization, we have tried various approaches but this described in the paper with Power BI was the most efficient. This paper aims to provide use case of effective implementation of Power BI tools in banking, more specifically in Credit Registry database, using the methodology of Big Data analytics and the features of Power BI tool.**

Keywords—Big Data, Power BI, Credit Registry

I. Introduction

With the exponential increasing of the amount of data, raises the need to understand trends in business and to gain important insights from the existing data. Different businesses need to understand analytical concepts using statistical methods, data prediction and machine learning [1]. These operations in the past were done just by developers, but now, these modern tools provide these abilities directly to people from the business. Microsoft Power BI is a tool for advanced analytics which enables normal users to use it and to leverage its capabilities for extracting useful knowledge from data, data visualizations and integration with R. Power BI empowers predictive analyzes by using machine learning without any previous programming.

Credit Registry is dataset that captures and persists financial information about credit borrowers and persists their history to contribute in improving the quality of loans and maintain the stability of banking system. This is one of the biggest datasets in central banks and one of the favorites for Big Data in central banks [2] [3]. In our case we have subset of the Credit Risk dataset of the Republic of North Macedonia. The data in the dataset is submitted by banks and saving houses and contains credit exposure data for the purposes of credit risk management. The Credit Registry is a collection of personal data, controlled by the central banks, which in our use case is anonymized and we work with extracted subset of the dataset [4].

This paper is organized as follows. In Section 2 we overview the Power BI features and our decision to use it. Then in Section 3 we describe the methodology and the process of performing the analysis workflow starting from the data source and ending with visualization. In Section 4 we describe our model created in Power BI Desktop, then in Section 5 analytics done on Power BI Service. The last Section reviews the benefits and future work which needs to be done.

II. Power BI

Power BI [5] is a business analytics tool that provides insights to enable fast and informed decisions. Power BI is a set of software services, applications, and connectors that work together to turn unrelated data sources into coherent, visual, and interactive displays. Data can be an Excel spreadsheet or a cloud-based hybrid data warehouse or local database. Power BI makes it easy to connect to data sources, visualize and discover important information, and share it through the web and mobile applications.

After reviewing multiple tools, we decided to use Power BI because of the high optimization and speed of data manipulation and analysis. In our example the 13GB database in SQL on import has been reduced to 330mb in Power BI .pbix format, since it has its own format that is adapted to handle big data [6]. Power BI supports around 115 types of data sources [7].

Our decision was done also according to Gartner [8], where Power BI is an application leader for business intelligence that has quickly become the most well-known and valuable to large corporations, surpassing Click and Tableau.

Power BI consists of:

- **Power BI Desktop** – Application for Windows desktop
- **Power BI Service** – Internet SaaS application named Power BI Service
- **Power BI Mobile** – Mobile application of Power Bi which works Windows, iOS and Android.

Power BI integrates with on-premises or cloud data sources, supporting all the data sources of Azure and many others.

When the data source changes not in a real-time manner, it is feasible to import the data in Power BI and then manipulate with top performances and easily analyze and visualize.

Large datasets which are more real-time data, can be leveraged by using Direct Query which get only needed data from the data source and efficiently display and refresh avoiding performance impacts.

Our data for this project is a subset of Big Data dataset which contains relational data, where the extraction is done in SQL Server.

Processing with Power BI starts with using Power BI Desktop when importing data, setting up relationships and visualization. Once the analysis is completed locally, it can be published on the SaaS version of Power BI on the cloud, where users can be granted access through Azure Active Directory user accounts. At the same time when the dataset, model and reports are published on SaaS, data, reports and displays are also available through the official mobile application. For a native mobile display, you need to edit your mobile layout via the desktop application.

Using this approach one can quickly and efficiently view and analyze analytics via desktop, web and mobile applications. Power BI can integrate reports into an external application through program code. To be able to embed external applications you need to have a Power BI Pro account with Azure Active Directory. The application needs to be registered in Azure Active Directory before it can access the Power BI REST API. Registering an application allows you to submit application identities. Power BI applications and their connections is shown in Fig. 1

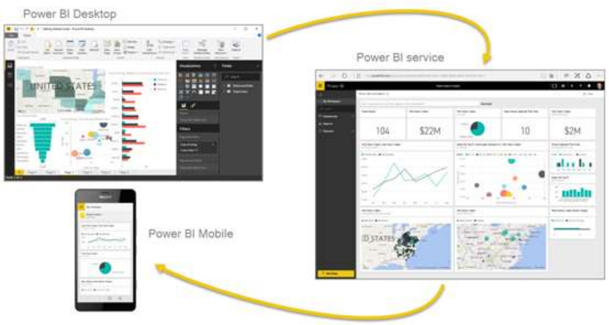


Fig. 1. Power BI applications and their connection

III. Process of performing the analysis

For performing the analysis, we have performed several stages of data processing to achieve the final results. Initially the data is imported from multiple sources, edited and updated in Power BI Desktop.

After creating models, reports and dashboards they are finally published and made available on the web and mobile application.

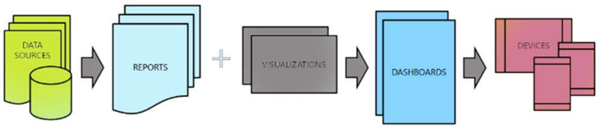


Fig. 2. Process of performing the analyses

The following is a detailed process from the data source to the final analysis.

A. *SQL Server process*

- **Analysis** - the SQL data source originally had more than 50 attributes. The minimum and maximum values are analyzed. It is important to note that due to legal changes over the years, not all fields have data.
- **Missing values** [9] - Missing values are initially located through SQL code, those that are few in a column are filled with the mean of the others for the corresponding attribute.
- **Feature selection** [10] - Locating the most important attributes is done with the help of fellow economists who are specialized in credit registry management. With their help, about 15 most important attributes are logged.
- **Feature engineering** – For improving the model efficiency we have derived new attributes from existing data [11].
    - **SizeOfBank** - derived column according to bank size code, which represents the size of the bank as a factor of three values.
    - **NumberOfLoans** – for every person using the SQL code we derived the number of loans in the current reporting period.
    - **DateStartLoan** - It has been found that the column for the first cash flow date has a rather illogical date due to the prior lack of control for that column, so a new column has been drawn that sets the credit start date at the time that credit party first appears in the database for the respective client.
    - **LoanDurationYears** - Derived column using the loan due date column.
    - **Age** - Age is derived for individuals through their identification number.

B. *Excel process - create additional data sources*

To improve the model and to set up and analyze with external factors, several sets of public data, which are an important economic factor, have been downloaded from the Internet.

- USD, EUR exchange rate list, to see how it affects the number of loans and their exposure.
- GDP data expressed in dollars. The purpose is to see how GDP is linked to the issued loans.
- Table of the municipalities of the Republic of North Macedonia, which has the number of municipalities required for merging and displaying the municipality's name for the loans.

C. *Power BI Desktop process*

- **Importing Data from SQL & Excel** - The data is imported into Power BI, a process that initially takes up to 20 minutes. The imported data is no longer dependent on the SQL and Excel sources. When changing data sources, Power BI only refreshes the data and retrieves it from scratch.
- **Creating Calculations** - Calculations are derived columns that are additionally calculated at the level of each row.
- **Creating Measures** - Measures are amounts at the table level, not at the level of each row. To create new information from data we used Data Analysis Expressions (DAX) which is a set of functions,

operators and constants that can be used in a formula or expression.

- **Creating hierarchies** - To enable drill down functionality, hierarchies of attributes have been created.
- **Model Design and Relationships** - In order all columns to behave as if they were in the same table, the model was created by merging all source tables through relationships.
- **Create Visualization Reports** - Reports are created using ready-made visualizations, then the corresponding columns displayed during the analysis are configured. For visualizing numeric attributes outliers, we installed additional extension from the AppSource to plot outliers with Box Plot.
- **Forecasting** - Power BI has built-in analysis using Linear Regression.
- **Creating views for mobile** - For neat display of mobile devices, the views for viewing from mobile phones are also designed.
- **Publishing Power BI Service** - After completing all the analysis and modeling locally, the dataset along with all visualizations are published and uploaded on the online cloud version of Power BI named Power BI Service.

D. Power BI Service process

- **Dashboard** - By selecting important visuals and displaying them on a new page or adding most important ones to dashboards.
- **Get insights** - Power BI Service has the ability to search on its own to learn new knowledge and visualizations of data. This knowledge is not always helpful. The Quick Insights feature automatically analyzes all the relations of data applying sophisticated algorithms to provide insights.
- **Ask question** – This is an option that allows writing queries in English and strives to produce results according to source columns and data.
- **Web publishing** - Ability to place graphs on external sites, but such access will also require external sites to have the user sign in details.

E. Power BI Mobile process

- **Mobile application testing** - You need to install the native Power BI Mobile application, and by logging into your Power BI account, all visualizations and views are displayed appropriately for mobile view.

IV. Analysis and Processing with Power BI Desktop

The full analysis is done with Power BI Desktop, the main application of the set of the Power BI tools. An important part is pre-processing process and then designing the model. Our model which is kind of a Star schema (see Fig. 3) and all the source charts are merged with the main credit registry dataset [12]. For having time series and grouping by months, quartiles and years we have used additional excel sheet which is imported and linked in the model.

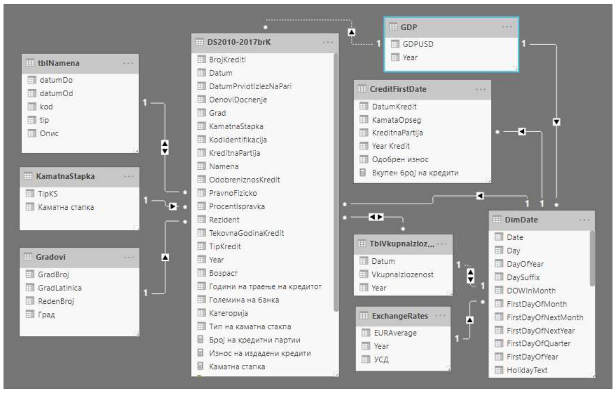
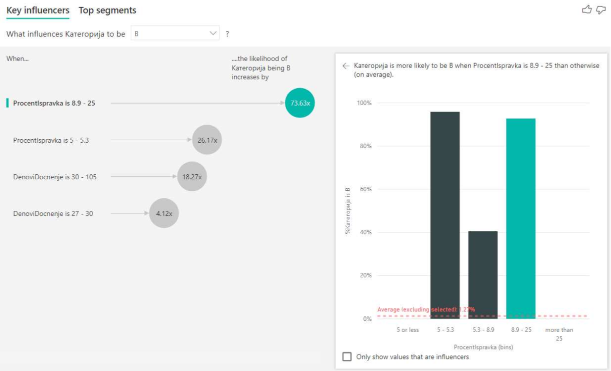


Fig. 3. Design of the model

After designing the model and creating new calculations and measures, the next step is to place visualizations on each side (report) and adjust them separately. It is beneficial that after configuring all visualizations separately, when selecting some information in one visualization (chart) it is updated in all visualizations in the level of report for the selected feature or range. Also for some reports with time series, is enabled and drilldown functionality.

The tool is powerful for finding dependencies between different attributes [13]. The following is an analysis of how the risk category depends on the day-to-day installment payment delay, and on the percentage of impairment entered by the banks themselves. The tool through Influencer visualization finds with high accuracy the range of days in which the client should be for every risk category, and also with more accuracy it detects the dependency for percentage of impairment. Figure 4 displays example that category is more likely to be B when the percentage of impairment is 8.9-25 with accuracy of 73.63%, and also when day-to-day installment payment delay is between 30-105 days.


Key influencers
Top segments
What influences Категорија to be
B
?
When...
....the likelihood of Категорија being B increases by
ProcentIspravka is 8.9 - 25
73.63x
ProcentIspravka is 5 - 5.3
26.17x
DenoviDocnenje is 30 - 105
18.27x
DenoviDocnenje is 27 - 30
4.12x
100%
80%
60%
40%
20%
0%
%Категорија is B
Average (excluding selected): 1.27%
5 or less
5 - 5.3
5.3 - 8.9
8.9 - 25
more than 25
ProcentIspravka (bins)
□ Only show values that are influencers

Категорија is more likely to be B when ProcentIspravka is 8.9 - 25 than otherwise (on average).

Fig. 4. Visualizing category dependencies through Influencer visualization

Business leaders can easily understand and work with data when they are in visualized. You can provide available visualizations in Power BI gallery or add custom visualizations including and R custom visualizations.

V. Analytics on Power BI Service

The Power BI Service enables full data upload and visualization of the Power BI cloud and its availability everywhere and at all times.

In the following there are some Figures of the solution and their description. Figure 4 shows analysis by age, number of loans by age of people and distribution of amount per age.

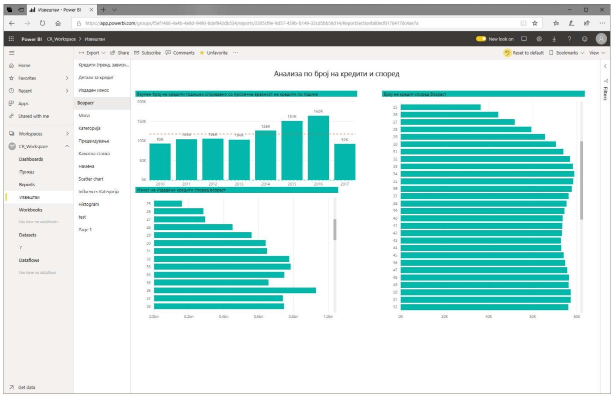


Fig. 5. Visualization by age, number of loans by age, and amount of loan by age

Figure 6 shows the distribution of number of loans per municipality of the country. The map is by Bing maps and using Power BI are displayed the sum of number of loans using green circles with appropriate size.

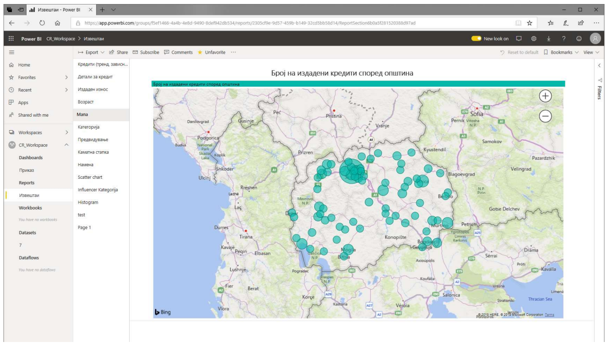


Fig. 6. Map of the municipalities in the Republic of Macedonia by number of loans

Figure 7 shows predicted values for period 2018-2020 using the data from previous layers. The prediction is integrated in Power BI and uses linear regression, which can be additionally configured.

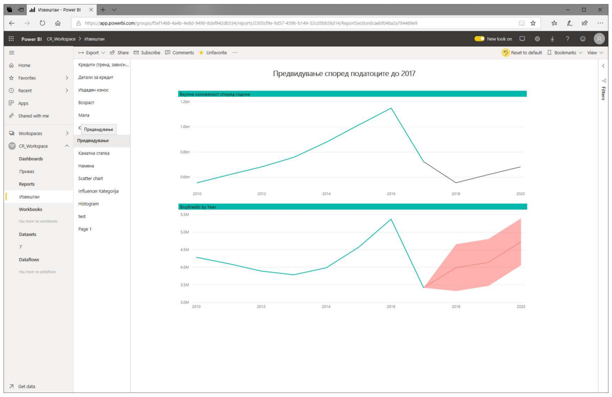

Извештаи - Power BI
https://app.powerbi.com/groups/f5ef1466-4a4b-4e8d-9490-8def942db534/reports/2305cf9e-9d57-459b-b149-32cd5bb58d14/ReportSection0caebf046a2a784469e9
Power BI CR_Workspace > Извештаи
Предвидување
Предвидување според податоците до 2017
Вкупна изложеност според година
1.2bn
1.0bn
0.8bn
0.6bn
2010
2012
2014
2016
2018
2020
BrojKrediti by Year
5.5M
5.0M
4.5M
4.0M
3.5M
3.0M
2010
2012
2014
2016
2018
2020

Fig. 7. Forecast visualization using Power BI feature, total exposure and number of credits per year.

VI. Conclusion

Power BI business analytics makes easier the analytics and visualization for all users of the company, gaining new insights and making effective advanced analytics. Using Power BI tool for Big Data Analytics is an effective way of discovering and visualizing knowledge in a very fast way, which unlike traditional tools is incomparably fast and powerful with modern capabilities for visualization, dependency and prediction. Dashboards in Power BI can transform the way people guide the business by supporting monitoring of social media, video streaming and real-time data.

With the help of Power BI one can achieve detailed insight and dependency analysis of the attributes themselves and correlation with external factors.

The project helped a lot to gain complete knowledge to the big data on the credit registry. The next step with this datasheet is to implement state-of-the-art machine learning algorithms for predicting and evaluating the results.

VII. References

[1] B. a. P. Z. Fang, "Big data in finance," *Big data concepts, theories, and applications.,* no. Springer, Cham, pp. 391-412, 2016.

[2] BearingPoint, "Big data in central banks: 2017 survey," 2017.

[3] C. B. M. P. J. L. &. S. Altavilla, "Banking supervision, monetary policy and risk-taking: Big data evidence from 15 credit registers.," 2020.

[4] Microsoft, "Advanced Analytics with Power BI White Paper,"

[5] A. Aspin, Pro Power BI Desktop, Apress, 2016.

[6] S. M. Ali, "Big data visualization: Tools and challenges," in *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016.

[7] "Power BI data sources," Microsoft, 10 03 2020. [Online]. Available: https://docs.microsoft.com/en-us/power-bi/power-bi-data-sources. [Accessed 13 03 2020].

[8] Microsoft, "Microsoft Power BI Blog," [Online]. Available: https://powerbi.microsoft.com/en-us/blog/microsoft-named-a-leader-in-gartners-2020-magic-quadrant-for-analytics-and-bi-platforms/. [Accessed 09 03 2020].

[9] A. C. Acock, "Working with missing values," *Journal of Marriage and family,* vol. 67.4, pp. 1012-1028, 2005.

[10] J. Li, "Feature selection: A data perspective," *ACM Computing Surveys* , vol. 50(6), pp. 1-45.

[11] M. R. a. M. C. Anderson, "Input selection for fast feature engineering," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 2016.

[12] P. a. P. G. Cerchiello, "Big data analysis for financial risk management," *Journal of Big Data (2016): ,* vol. 3.1, 2016.

[13] C. S. C. a. L. G. Cantú, "How Do Bank-Specific Characteristics Affect Lending? New Evidence Based on Credit Registry Data from Latin America," in *Bank for International Settlements (BIS)*, 2019.