Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 7001-7007
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.3515

Leveraging data quality analytics for enhanced loan application and performance insights: A comprehensive approach

Brunela Karamani[1*], Bojken Shehu[1]
[1,2]Department of Computer Science,Faculty of Information Technology, University Polytechnic of Tirana, Albania; bkaramani@fti.edu.al (B.K.) bshehu@fti.edu.al (B.S.)

Abstract: The objective of this research is to discover the ways in which the banking sector can use data quality analytics to improve the processing of applications for loans and gain better insights into their performance. We use SQL, Power BI, and Python to examine how these decisions can be improved, how customer experiences can be made better and more compliant, and how we can use good data analytics to ensure that proper regulatory guidelines are followed. To do this, we employ a quasi-mixed method that solves the "what" question with descriptive and inferential statistics, the "how" question with good data visualization, and the "why" question using advanced analytics. The data we use encompass over 300K records of loan performance and over 150K entries of loan applications from the years 2020–2022. Data cleaning, transformation, analysis, and visualization are done using SQL, Python, and Power BI to derive actionable insights. The analytics side of data quality is examined in this study with a view to understanding its impact on the loan application process and the performance of the business. The analytics have shown that the quality of data and the insight which can be gained from it has a material effect on the aforementioned processes and performance. The study shows how sophisticated analysis can be applied in the actual operations of a banking institution. It underscores the value of solid customer and loan data for making high-level decisions within the bank.

Keywords: *Banking, Customer satisfaction, Data analytics, Data management, Data quality, Loan application, Loan performance, Performance analysis.*

1. Introduction

In the contemporary landscape of data-driven banking, the integrity and comprehensiveness of customer data are paramount for effective credit rating assessments and the analysis of loan performance [1]. The intensifying competition resulting from evolving consumer preferences and the liberalization of the banking sector necessitates that financial institutions adopt data-centric strategies in their decision-making processes [2]. This data-driven management not only facilitates the delivery of personalized banking solutions, thereby enhancing customer experiences but also assists in meeting regulatory requirements and conducting thorough internal audits [3].

High-quality customer data serves as a foundational element for deriving actionable insights, particularly concerning the dynamics of consumer behavior concerning loan applications and repayment success [6]. By analyzing various dimensions such as demographics, transaction histories, and engagement patterns, banks can identify emerging trends and preferences, thus enabling the customization of their loan products and services to align with the needs of economically active clients [4].

The relationship between personalization and customer expectations in banking has become increasingly pronounced, with clients demanding a tailored approach in their interactions and offerings—particularly in the context of loans [5]. Banks leveraging data quality analytics can precisely segment their markets, discern individual preferences, and anticipate future loan requirements [7]. This capability allows for the formulation of targeted marketing campaigns, the recommendation of bespoke

* Correspondence: bkaramani@fti.edu.al

loan products, and the provision of enhanced customer support, thereby granting banks a competitive advantage in the dynamic loan market and fostering customer loyalty [8].

In an era characterized by an exponential increase in data volume, ensuring accuracy and reliability in customer data is critical, particularly in the context of loan-related information. Missteps or inaccuracies can lead to detrimental outcomes, underscoring the necessity for financial institutions to prioritize data quality analytics projects focused on data cleansing, standardization, and the maintenance of dataset integrity [9].

The advantages of data quality analytics in banking extend, particularly to loan application processing and performance solutions, encompassing facets such as efficient loan origination, regulatory compliance, and enhanced cost-effectiveness [10]. The significance of high-quality client data is further emphasized by its role in delivering credit-related benefits and securing a competitive edge in the financial services sector. Consequently, advanced analytics techniques and robust data governance policies are essential for banks aiming to enhance operational efficiency, maintain regulatory compliance, and pursue sustainable growth in loan services.

2. Related Work

The research presented in article [9] introduces the I-BiDaaS platform, which provides an innovative and user-friendly solution for managing big data analytics in the banking sector, allowing employees without extensive data science expertise to engage in meaningful analytics. This work addresses the inherent tension between data privacy regulations and the necessity for diverse data utilization, presenting examples of effective cybersecurity measures and customer protection strategies.

Further, the findings referenced in [10] utilize actual case studies from the banking sector to elucidate the implementation of big data analytics in marketing strategies. This research identifies critical non-technical success factors necessary for the effective adoption of big data, including the establishment of deeper analytics, an appropriate attribution baseline, and alignment of organizational goals.

The challenges associated with managing large datasets are acknowledged, with an emphasis on the need for results that are user-friendly and easily interpretable. The application of dimension reduction techniques is deemed essential for facilitating practical decision-support applications [2].

Additionally, report [1] highlights the Oracle EDQ platform's efficacy in addressing the challenges of data integration and transformation, thereby ensuring an efficient analytical process. This research underscores the importance of data cleaning and preprocessing in deriving insightful interpretations from customer data.

Finally, the study by Li and Beloglazov [7] discusses the application of data analytics in smart manufacturing, identifying research gaps in data quality management and statistical process control, and serving as a foundation for future academic inquiry into big data applications.

3. Methodology

The methodology framework developed for this study is designed to systematically enrich customer data for loan determination and performance assessment within the banking sector. This framework employs both descriptive and inferential statistics to illustrate how data quality analysis techniques can enhance the customer experience concerning loan-relevant information.

The methodology commences with a robust application of descriptive statistics, aimed at analyzing features and patterns within borrower data derived from loan applications and their respective performance. Various statistical measures, including mean, median, standard deviation, and frequency distribution, are utilized to profile the loan-related data, thereby elucidating its distribution and variability. This process facilitates the identification of significant patterns, detection of masking data, and clarification of the primary characteristics underlying loan processes. Following the application of descriptive statistics, the methodology advances to inferential statistics, which allows for a deeper exploration of loan-related data, leading to broader theoretical implications that inform business decisions. Inferential statistics enable a comprehensive understanding of the relationships among variables, facilitating connections with relevant parameters. Techniques such as hypothesis testing,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 7001-7007, 2024
DOI: 10.55214/25768484.v8i6.3515

confidence intervals, and regression analysis are employed to reveal perceived relationships, form future projections, and generate definitive interpretations of patterns observed in loan applications and their associated metrics.

Through rigorous data science analysis of loan-related information, the institution ensures both quality and timeliness by implementing specialized analytical methods for evaluating all loan applications and performance insights. This approach enhances the accuracy and reliability of creditworthiness assessments, thereby improving the management of consumer data.

Subsequently, data cleaning methods are applied to address missing values, outliers, and inconsistencies, which are essential steps before data standardization processes that ensure uniformity and consistency across the dataset. Advanced analytical techniques, including machine learning algorithms, are also employed to identify complex relationships within the data and enhance loan-related procedures.

Furthermore, the methodology incorporates a stringent verification process to ascertain the truthfulness and accuracy of data interpretations. External validation procedures, utilizing cross-validation methods, are implemented to confirm the model's stability and generalizability. Regression diagnostics, such as sensitivity analyses, assess the robustness of conclusions drawn from the model results.

This research process necessitates extensive collaboration with field experts and stakeholders within the banking sector, whose feedback contributes to the refinement of weighting models. Such collaboration aids practitioners in focusing on pertinent decision-making processes related to default risk and credit scoring, ultimately enhancing the effectiveness and applicability of the findings.

4. Results and Discussions

The analysis conducted in this study delved into the intricate realm of data quality analytics within an Albanian bank, specifically focusing on loan application and performance insights. Through meticulous examination grounded in descriptive and inferential statistics, the study aimed to refine customer data to extract actionable insights crucial for strategic decision-making. The data that has been observed during the period 2020 - 2022. We have taken into consideration the customer data 100K, the loan application data 150K, and the loan performance aggregated files 300K. To undertake data analytics, we employ SQL, Python, and Power BI, following these steps:

a) Data Import and Preparation: We get customer data, loan application data, and loan performance data from the database to our SQL database. Through SND queries, we create and process data to eliminate null values, outliers, and abnormal data.

b) Data Analysis with SQL: The first step we take is writing SQL queries that allow us to conduct descriptive statistics analysis (EDA) on each data set separately during our integration phase. We compute mean, median, mode, and standard deviation values from numerical variables and present the descriptive statistics for our data. Besides we perform inferences statistics as, for instance, hypothesis testing or correlation analysis, which are aimed at discovering the correlations between factors.

c) Data Integration and Feature Engineering: Customer data and loan application data are combined by joining them by SQL join method or merge operations, subsequently followed by performance of customer loan data. We perform feature engineering termed as the creation of new variables or the transformation of the given integrated dataset into meaningful insights.

d) Data Analytics with Python: We rely on the use of built-in Python Libraries Pandas and NumPy for deep data analyses and modeling. We use highly advanced analytics capabilities, such as predictive modeling and clustering analysis, which help us to understand better loan application information and performance.

e) Integration with Power BI: The Data we download is in the clean and analyzed form, so we directly bring it to the Power BI Desktop for analysis. Through Power BI's SQL Server connector, we have our database directly linked with it to import data into our SQL server.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 7001-7007, 2024
DOI: 10.55214/25768484.v8i6.3515

f) Strategic Decision-Making: We utilize the information received from the analytics to make reasonable decisions about the criterion for loan approval, risk analysis, segmentation of customers, and so on. We convey the findings and recommendations to the stakeholders.

Table 1.
The observation data by product base and customer type.

Observation period 2020 - 2022	Products base		
	Credit card	Loan	Overdraft
Company	4.376	9.933	1.749
Individual	39.785	85.872	8.285
Total number	44.162	95.805	10.034
Total %	29%	64%	7%

In total, 55 variables were taken into consideration, 12 from customer data (data set 1), 25 from loan application data (data set 2) and 18 from loan performance data (data set 3), which were analyzed and classified according to their quality as good, average, poor. Overall, most of the data points studied have good to acceptable data quality.

Table 2.
The data point per data set with quality

Quality	Good	Average	Poor
data set 1	10	1	1
data set 2	22	3	0
data set 3	17	1	0
Total	49	5	1

The analysis of roll rates is useful for identifying the appropriate definition of a "bad" loan product. On the other hand, the max delinquency analysis helps to refine the sampling process and offers valuable insights into application volumes. The variables are shown below in order of quality and missing value percentage (minimum and maximum) based on the data type.

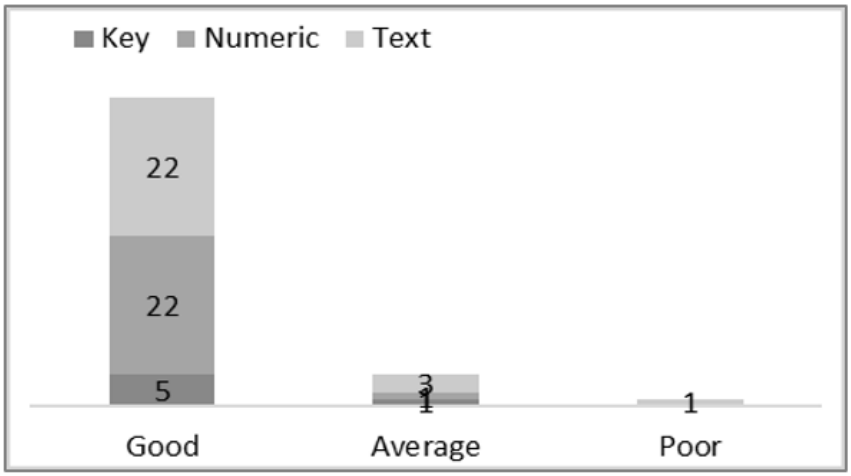

Figure 1.
Quality of data type.

In data set 2 after analyzing the data results: 80% of customers have 1 record, 15% 2 records and 5% 3+ records. The rejection rate for the observed period is 4.6%. The average amount applied in ALL is 242K with average tenor 27 months and with average monthly income in ALL 46k.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 7001-7007, 2024
DOI: 10.55214/25768484.v8i6.3515

For data set 3 the results after the data analysis are: 90% of customer are individual. 10% of loans are restructured and 5.6% of loans are in PAR (In arrears above 30 days).

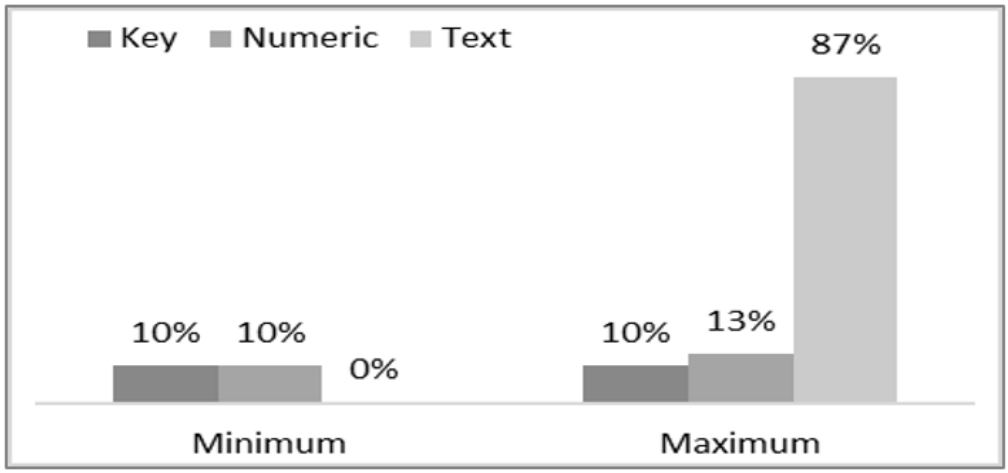

Figure 2.
Missing data in %.

The analysis of roll rates is useful for identifying the appropriate definition of a "bad" loan product. On the other hand, the max delinquency analysis helps to refine the sampling process and offers valuable insights into application volumes.

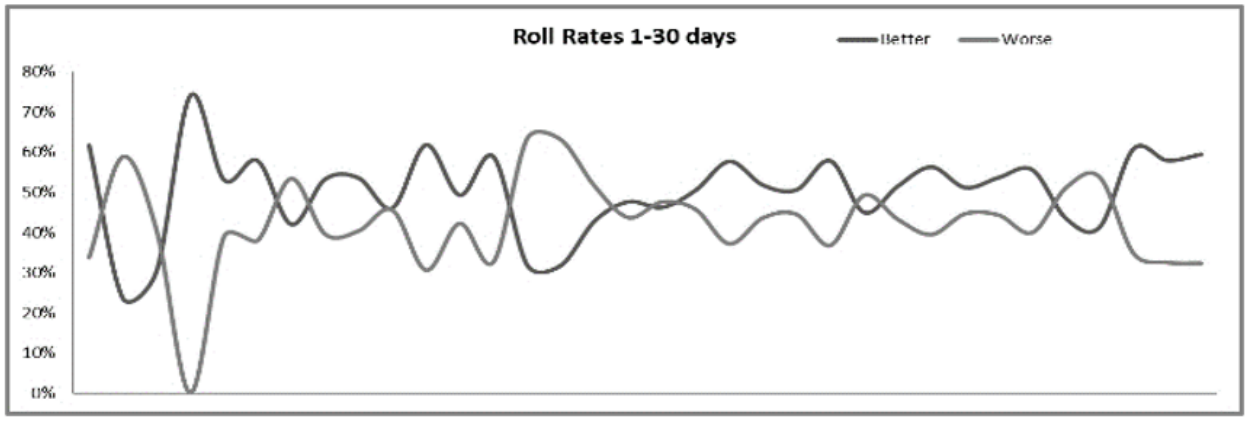

Figure 3.
Delinquency Status (Better; Worse) Roll Rates 1- 30 days.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 7001-7007, 2024
DOI: 10.55214/25768484.v8i6.3515

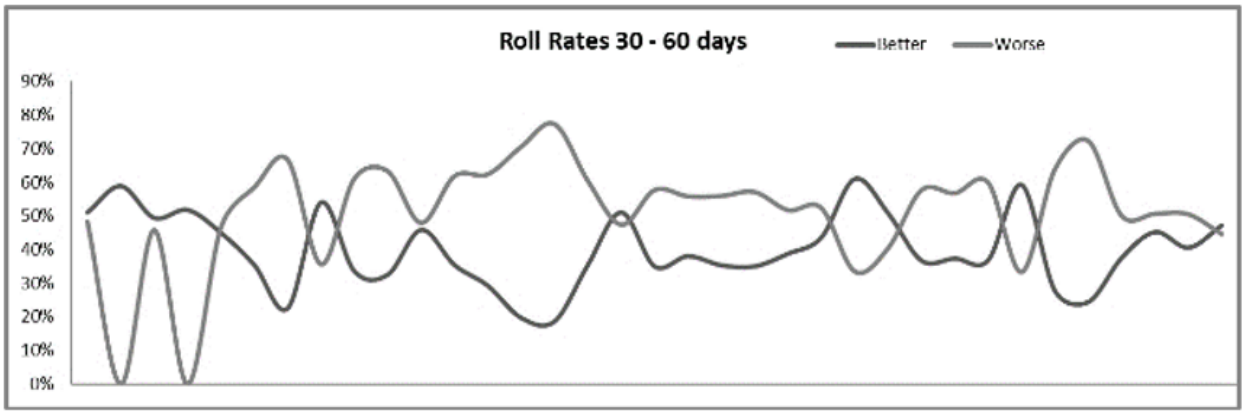

Figure 4.
Delinquency Status (Better; Worse) Roll Rates 30 - 60 days.

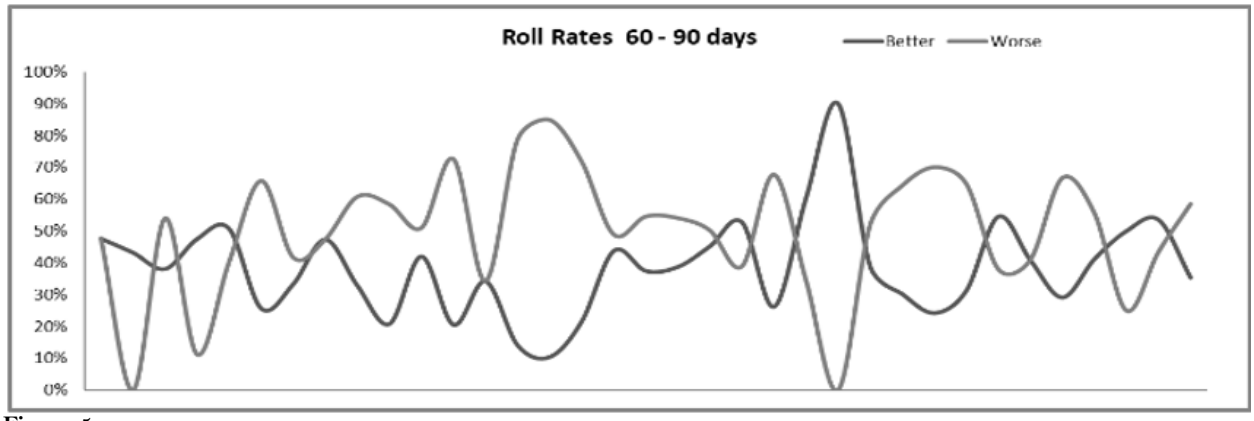

Figure 5.
Delinquency Status (Better; Worse) Roll Rates 60 - 90 days.

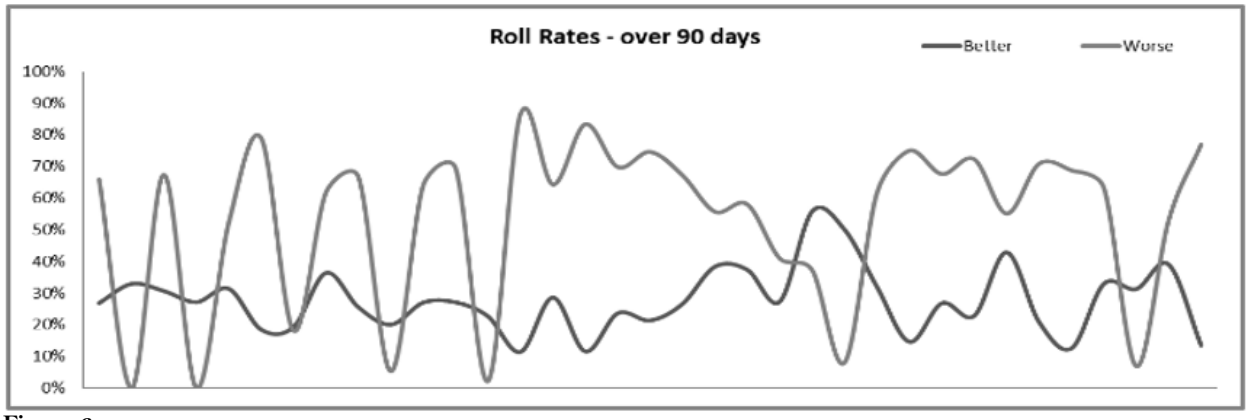

Figure 6.
Delinquency Status (Better; Worse) Roll Rates over 90 days.

The focus of the paper is to highlight the crucial nature of data quality in the banking industry, having customer datasets in mind. First of all, it is important to put the right data quality processes and tools in place to get data accuracy, regulate compliance, and gain a competitive advantage. This goal can be accomplished by blending SQL for data formatting and initial analysis, applying Python to the more sophisticated analysis and modeling, and Power BI as the tool for visualization and reporting. The

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 7001-7007, 2024
DOI: 10.55214/25768484.v8i6.3515

research will be guided by a comprehensive regulation that focuses on giving banks crucial insights for enhancing loan application systems, improving loan performance, and thereby encouraging strategic growth and innovation in the competitive banking industry.

5. Conclusions

This research elucidates the substantial significance of data quality analytics in augmenting the loan application procedure and performance evaluations within the banking domain. By assimilating instruments such as SQL, Power BI, and Python, financial institutions can exploit customer and loan data to render judicious decisions, enhance operational efficacy, and propel strategic expansion. The results underscore the paramount importance of guaranteeing high-quality data in banking, as it empowers institutions to more effectively evaluate creditworthiness, mitigate loan default risks, and provide customized financial products to satisfy customer requirements.

The application of SQL for data cleansing and standardization, Python for sophisticated statistical analysis, and Power BI for real-time data visualization proved exceptionally efficacious in converting extensive datasets into pragmatic insights. The investigation revealed that pivotal performance indicators, such as customer loan default rates and repayment conduct, could be markedly improved through this all-encompassing methodology. Furthermore, the utilization of predictive modeling and clustering methodologies facilitated more precise customer segmentation, culminating in enhanced loan product recommendations and a diminution in credit risk.

In conclusion, this manuscript accentuates the pragmatic ramifications of harnessing advanced analytics and BI instruments in the banking sector. It emphasizes the imperative for financial institutions to allocate resources to data quality initiatives and embrace comprehensive analytical frameworks that can underpin improved decision-making, customer segmentation, and regulatory adherence. The amalgamation of SQL, Power BI, and Python has demonstrated to be a remarkably effective synthesis for enabling banks to realize these objectives, ultimately fostering operational enhancements and promoting sustainable growth within the dynamic banking industry.

References

[1] Karamani, B. Data Quality Analytics for Customer-Level Dataset, 5. *2023 ICITEE The 2th International Conference on Information Technologies and Educational Engineering 15-16 December, Polytechnic University of Tirana, Albania*, 2023.

[2] Sharma, R. Analytics in Banking: A Comprehensive Guide to Risk, Marketing, Pricing, and Product Development. *Wiley*, 2019.

[3] Janssen, M., & Kuk, G. Governance and business process management in banking. *Business Process Management Journal*, 22(6), 1124-1148, 2016.

[4] Deloitte. Banking Personalization: 3 Keys to Increasing Customer Engagement, 2021. Retrieved from: *https://www2.deloitte.com/us/en/insights/industry/financial-services/personalization-banking-industry-customer-engagement.html*

[5] Mishra, D., & Mishra, A. Personalization in Banking Services: A Study of Indian Banks. *Journal of Digital Banking*, 4(3), 256-272,2020.

[6] Redman, T.C. Data Driven: Creating a Data Culture. *Harvard Business Review Press*. 2018

[7] Li, S., & Beloglazov, A. Data Quality Assessment Framework for Banking Systems. *IEEE Access*, 8, 84731-84744, 2020.

[8] Xing, H., Peng, M., & Zhu, H. A Comprehensive Review of Data Source Selection Methods for Data Analysis in Big Data Environment. *6th International Conference on Control, Automation and Robotics (ICCAR)* pp. 59-64, 2020.

[9] Alexopoulos, A. et al. Big Data Analytics in the Banking Sector: Guidelines and Lessons Learned from the CaixaBank Case, *Technologies, and Applications for Big Data Value* pp 273–297, 2022. https://link.springer.com/chapter/10.1007/978-3-030-78307-5_13

[10] He, W., Hung, J.-L. and Liu, L. Impact of Big Data Analytics on Banking: A Case Study, *Journal of Enterprise Information Management*, Vol. 36 No. 2, pp. 459-479. 2023. https://doi.org/10.1108/JEIM-05-2020-0176