European Research Studies Journal
Volume XXVII, Special Issue 2, 2024

pp. 83-93

A Graph-Based Recommendation System Leveraging Cosine Similarity for Enhanced Marketing Decisions

Submitted 18/02/24, 1st revision 16/03/24, 2nd revision 20/04/24, accepted 16/05/24

Tomasz Smutek[1], Marcin Kowalski[2], Olena Ivashko[3], Robert Chmura[4], Justyna Sokołowska-Woźniak[5]

Abstract:

Purpose: *This work aims to present a comprehensive customer recommendation system based on cosine similarity. The primary objective is to develop an effective tool that assists sellers in identifying and recommending similar customers by analyzing their characteristics and behaviors.*
Design/Methodology/Approach: *The methodology analyzes demographic data, purchase history, and other customer characteristics to calculate cosine similarity. This process includes data processing techniques such as feature integration and generating a cosine similarity matrix. The results demonstrate the system's effectiveness through thorough analysis.*
Findings: *The analysis confirms the effectiveness of the proposed recommendation system, revealing that using cosine similarity can identify and recommend similar customers accurately.*
Practical Implications: *The study emphasizes incorporating modern data analysis methods into marketing and customer relationship management. This approach can enhance the efficiency of sales activities and elevate customer satisfaction.*
Originality/Value: *This work offers a novel approach to customer recommendations by employing cosine similarity and innovative data processing techniques. It demonstrates how advanced data analysis methods can be leveraged to improve sales strategies and foster stronger customer relationships.*

Keywords: *Recommendation system, cosine similarity, customer behavior analysis.*

JEL codes: *C45, C63, D83, L81, M31.*

Paper type: *Research article.*

[1]*Corresponding Author: WSEI University, Lublin, Poland, e-mail: Tomasz.Smutek@wsei.lublin.pl;*
[2]*WSEI University, Lublin, Poland, e-mail: Marcin.Kowalski@wsei.lublin.pl;*
[3]*WSEI University, Lublin, Poland, e-mail: Olena.Ivashko@wsei.lublin.pl;*
[4]*WSEI University, Lublin, Poland, e-mail: Robert.Chmura@wsei.lublin.pl;*
[5]*Wyższa Szkoła Biznesu - National Louis University, e-mail: sokolowj@wsb-nlu.edu.pl;*

1. Introduction

A product recommendation system, also called a recommendation engine, is a tool that generates and delivers recommended products to a specific user based on their previous activity and preferences. This mechanism enables a closer understanding of consumers' habits and needs and allows for analyzing the behavior of particular users on store websites. Thanks to this, online stores can adapt their offer to individual customer preferences, which contributes to increased shopping satisfaction and sales.

The three main types of recommendation engines are collaborative filtering, content-based filtering, and complementary filtering. Collaborative filtering analyzes user behavior data to suggest products to other users with similar interests, shopping habits, and preferences. In this method, it is not necessary to understand the exact nature of the product because the algorithm is based on purchasing preferences, purchasing decisions, and the browsing history of the user's previous sessions. This makes it possible to create comprehensive product recommendations.

In turn, content-based filtering analyzes the product description and features, such as category, price, segment, technical specifications, and appearance. The algorithm looks for similarities between the specifications of two products and suggests other products that may interest the user based on these characteristics.

Complementary filtering, on the other hand, is a technique that analyzes the probability of purchasing several products simultaneously. Based on the purchase history of other users, the mechanism can determine which products are often purchased together, which allows the user to suggest adding additional products to their basket.

These three types of recommendation engines are widely used in various fields, such as e-commerce, streaming platforms, and social networking sites, to provide users with personalized and accurate product and content recommendations (Smith and Brown, 2023; Johnson and Patel, 2024; Wang and Li, 2023; Chen and Zhang, 2022; Gupta and Kumar, 2024; Zampeta and Chondrokoukis, 2023).

The method of indicating recommendations proposed by Recostream is an innovative way that allows the consumer to choose a set of recommendations to browse. Using a combination of the mentioned techniques - collaborative filtering, content-based filtering, and complementary filtering - this system simultaneously provides the user with four sets of recommendations. By placing a small window that perfectly matches the store's website, users can choose which recommendations they want to see at a given moment.

Thanks to this variety of recommendation models, the consumer not only has a wide range of personalized suggestions at his disposal but also the process of finding the

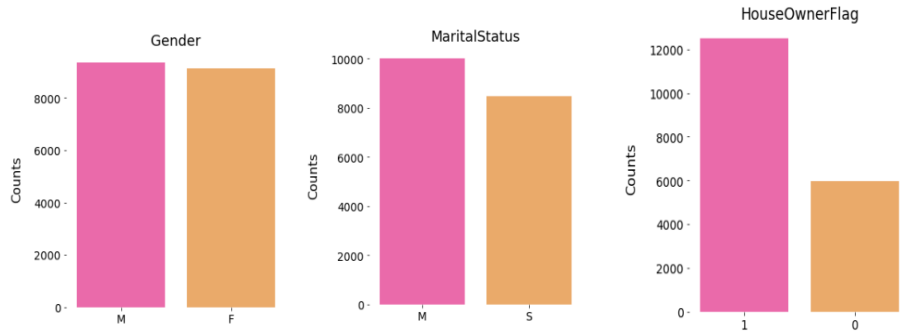
product he is interested in is significantly simplified, eliminating the need to browse the entire store website. This interactive way of presenting recommendations not only increases the attractiveness of using the store website but also improves the customer's shopping experience, positively influencing their satisfaction and willingness to return to the store in the future (Rymarczyk and Sikora, 2024; Kulisz *et al.*, 2024; Thalassinos *et al.*, 2013).

2. Sales Analysis

In the statistical analyses performed for our clients, we used a function providing basic descriptive statistics and distribution charts for demographic variables. The function provided the following statistics for int and float variables: variance, standard deviation, median, and mean. Additionally, distribution plots were generated for these variables.

However, for variables of a different type, the function provided the number of customers in a given category and generated appropriate charts. It is worth noting that we have 9,351 customers in total, of which 9,133 customers were included in the gender breakdown analysis. This similar number of customers by gender suggests that the sample is representative and allows for reliable inferences about the customer population (Figure 1).

Figure 1. *Division (a) by gender, (b) by marital status, (c) and by having your own home*



(a) (b) (c)

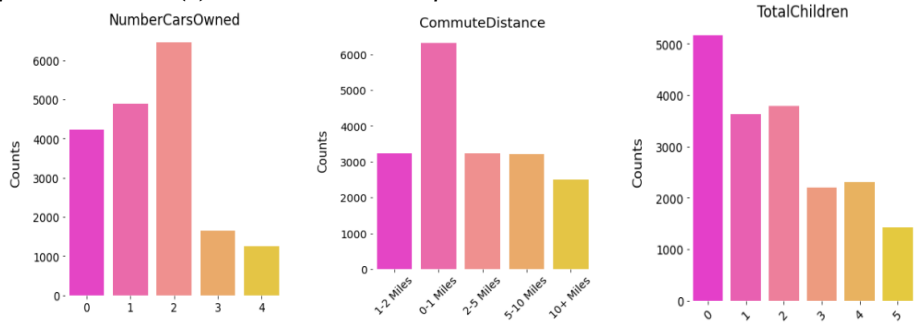
Source: *Own creation.*

Analysis of Figure 1(b) shows that there are slightly more customers who are married (marked with the letter M) than those who are single (marked with the letter S). However, from Figure 1(c), we can read that the number of customers who own their own homes is more than twice as large as the number of customers who do not own their own homes.

Chart Figure 2a shows the division of customers according to the number of cars they own. Among customers, most people own two vehicles (over 6,000 people).

A similar number of customers have 1 or 0 vehicles (between 4 and 5 thousand people). Much fewer customers have 3 or 4 cars).

Figure 2. *Division according to (a) the number of cars owned, (b) the distance from the store, (c) the total number of children*



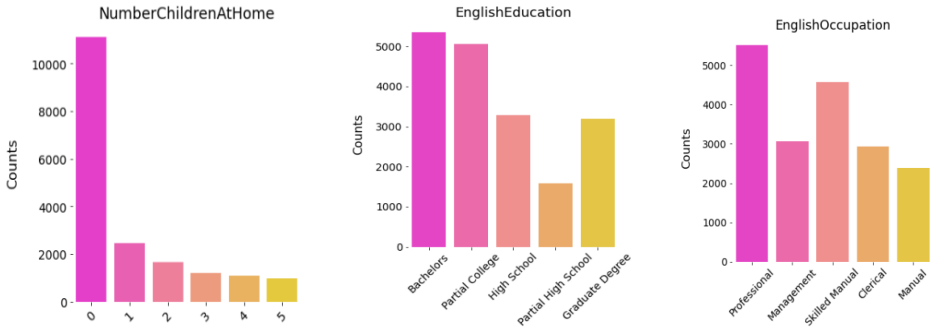
(a) (b) (c)

Source: *Own creation.*

Figure 2b shows the distance of customers from the store. Most customers live within 0 to 1 mile of the store, which is very close. This is over 6,000 customers). About 3,000 customers each live within 1-2 miles, 2-5 miles, and 5-10 miles of the store. The fewest customers live more than 10 miles from the store, but their number is almost 3,000.

Customers were also divided by the number of children they had. Most customers have at least one child (Figure 2c).

Figure 3*. Division according to (a) the number of children at home, (b) education, (c) type of profession*



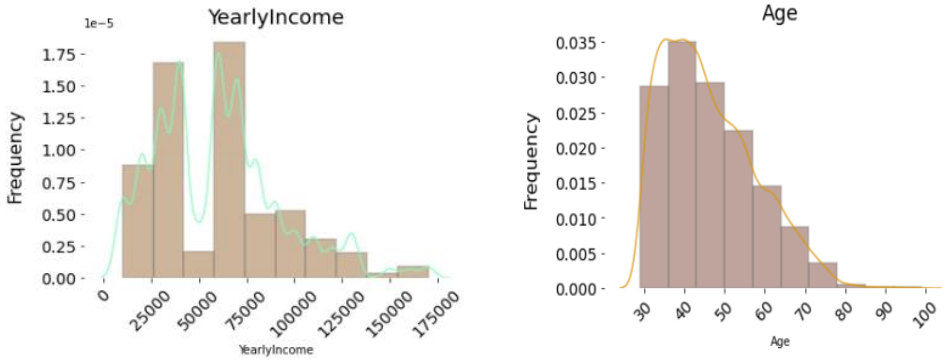
(a) (b) (c)

Source: *Own creation.*

In addition to the division of clients according to the number of children they have, we also analyzed the division according to the number of children at home (Figure 3a). Most clients do not have children.

Figure 3b shows that most customers have higher education. Few clients have not completed high school. Most customers have the profession type "Professional" followed by "Skilled manual." The "Manual" job type has the fewest clients (Figure 3c).

Figure 4. *Division by (a) annual income, (b) age*

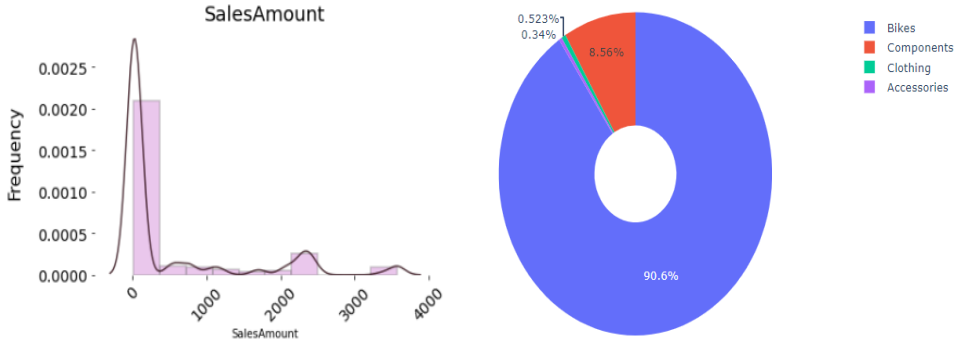


(a) (b)

Source: *Own creation.*

Figure 4a shows a histogram of customers' annual income. A yearly income often appeared between 25,000 and 50,000 and between 50,000 and 75,000. Based on the BirthDate column (Figure 4b), an age column was added.

Figure 5. *Sales amount (a) and a pie chart showing the percentage of sales amount divided into product categories (b)*



Source: *Own creation.*

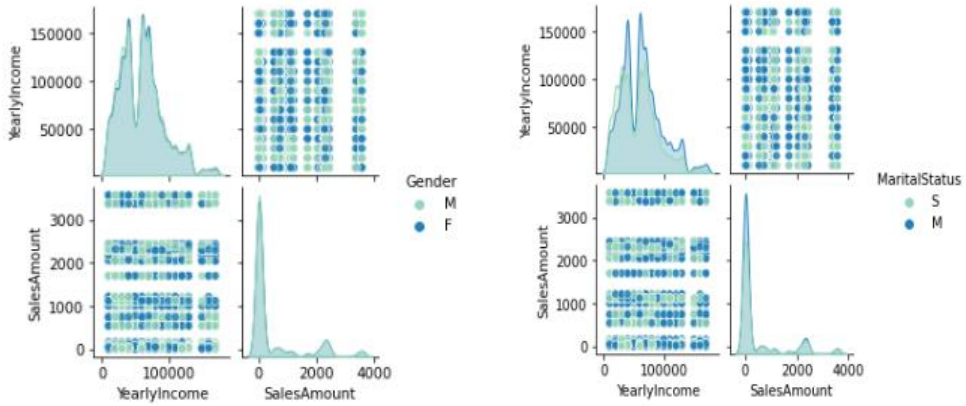
The distribution of the sales amount is asymmetric (Figure 5). The average sales amount is 486 (Figure 6).

2.1 Model of a Graph Recommendation System

The goal is to incorporate a recommendation system into the graph model that uses a matrix of distances between customers encoded using autoencoders. Customer data

extracted from the AdventureWorks database, as well as population and average GDP information for customer areas, were collected and encoded using an autoencoder.

Figure 6. *Sales amount and annual income by gender and marital status*



Source: *Own creation.*

Dimension reduction was then performed using principal component analysis (PCA). After coding customer data and reducing dimensions using PCA, two matrices were calculated: the Euclidean distance matrix between customers and the similarity matrix. The Euclidean distance matrix represents the distance between each pair of customers in a reduced dimension space. The similarity matrix was determined by subtracting the Euclidean distances from one, which allows for measuring customer similarity.

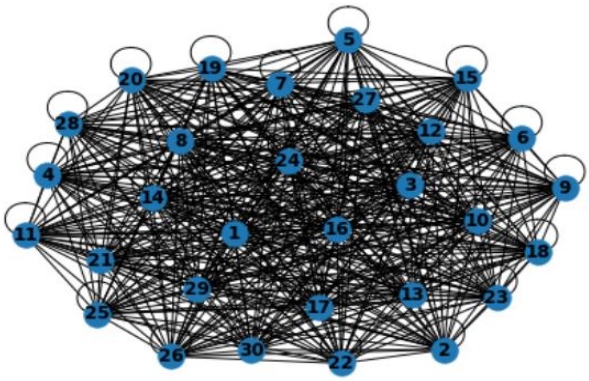
These customer distance or similarity matrices can now be incorporated into the graph model as the primary input. Using this data in a graph model will enable the identification of similar customer groups and recommendation of appropriate marketing activities for these groups.

Thanks to this recommendation system, the seller will understand customer preferences and needs better and adapt marketing strategies more effectively, which can help increase customer satisfaction and sales results.

A graph was created based on the similarity matrix and using the NetworkX library. We introduced a similarity matrix to the constructor and obtained N nodes (as many as we have observations) and NxN edges (representing the similarities between each pair of observations).

The edge weight reflects the similarity value (0 – dissimilar, 1 – 100% similar). A graph visualization was created for 30 selected clients (Figure 7). First, a graph was made using nx. Draw where each vertex is connected to the others.

Figure 7. *Graph drawing using nx. draw*



Source: *Own creation.*

In creating a graph visualization using Plotly, we first collect vertex positions, color, and text from an existing graph. This information is necessary to make a scatter plot in Plotly, where each vertex is represented as a point. Then, for each edge in the graph, we create a line graph that connects the two points representing the vertices that are connected by that edge.

In this way, we create edge traces that are added to the trace list and returned for later use. After making a graph visualization, we add interactions, such as an interactive slider. The slider allows the user to control the similarity values between customers included in the chart. For example, edges with a weight less than a certain similarity threshold may be hidden or displayed depending on the slider settings.

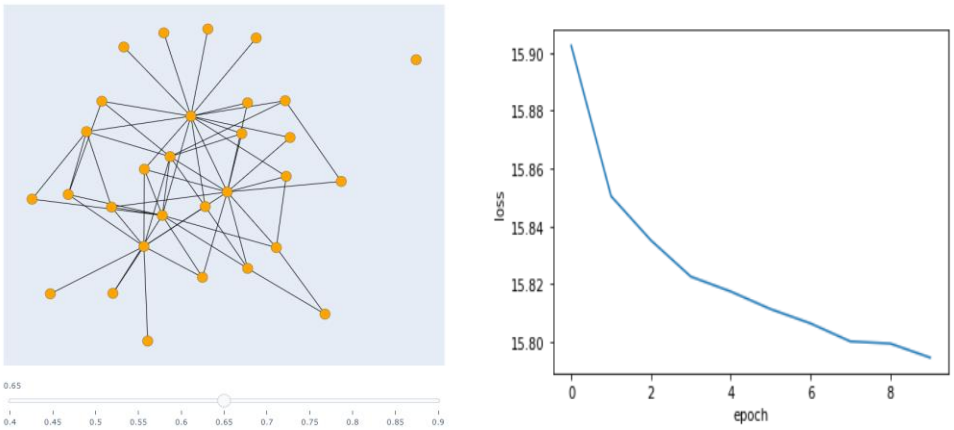
We add the appropriate customer key (CustomerKey) and the customer's e-mail address to each vertex, which is visible when you hover the mouse over a given vertex. This information is helpful for the user in identifying specific customers in the visualization.

As a result, we obtain an interactive graph visualization with an interactive slider that allows you to control the visibility of edges based on the similarity value between clients. For example, only customers with a similarity greater than or equal to 0.65 can be displayed in the figure, allowing for a better understanding and analysis of the graph structure (Figure 8).

In the next step, a model was built using the created graph. The test set was created through negative sampling, and then the model training began. The training parameters included ten epochs and a batch size of 256. The graph below shows the loss values in the test set during model training. We observe that the loss value decreases with subsequent training epochs, which indicates an improvement in the model's performance (Figure 9).

The recommendation system returned the most similar customer for the given customers, with the user able to control the number of returned customers. This process allows recommendations to be more precisely tailored to the user's needs by controlling the number of suggested customers.

Figure 8. *Graph visualization with an interactive slider (a) and Loss function (b)*



Source: *Own creation.*

The customer recommendation system is based on the use of cosine similarity between them, which allows the identification of similar customers based on their characteristics. This process involves several steps. First, you use a dataset of customer information that includes various features such as demographics, purchase history, etc. Then, you define a function that combines selected features from the appropriate rows of the dataset and returns one combined feature for each customer.

After creating the function, a new column is added to the dataset containing the function's results applied to each row. This column will contain the combined characteristics of each client. The next step is to calculate the cosine_sim matrix, a NumPy array with the calculated cosine similarity between each client. The values in the matrix indicate the degree of similarity between customers.

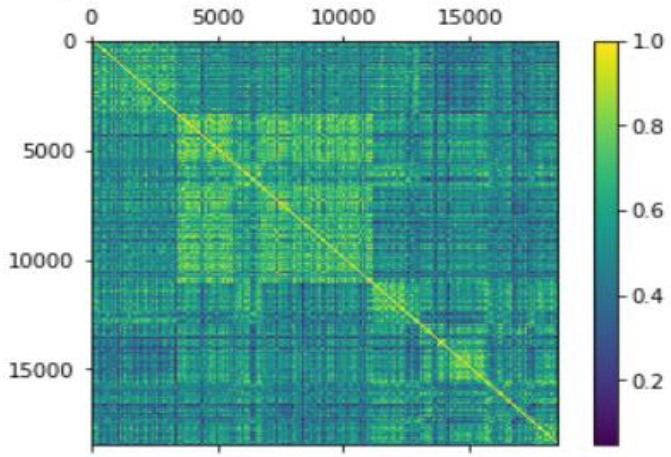
For example, if the value in the matrix is 1, these two customers are 100% similar (they are the same customer). The value 1 appears on the diagonal of the matrix because each customer is 100% identical to each other (Figure 9).

2.2 Methods for Detecting Communities in a Graph

Graph community detection is an essential branch of social network and graph analysis that deals with identifying groups in a graph where nodes are more connected to each other than to nodes outside the group.

Many methods are used for this purpose, each with its advantages and limitations, depending on the specifics of the graph and research goals. One popular method is the modularity method, which seeks to maximize modularity in a graph by dividing it into communities. Modularity-based algorithms try to find a graph partition that increases the number of edges within the community and minimizes the number of edges between communities.

Figure 9. *Matrix visualization with customer similarities*



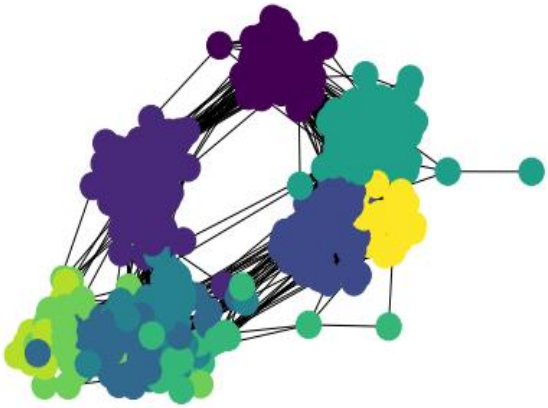
Source: *Own creation.*

Another popular technique is label propagation-based algorithms, which exchange community membership information between neighboring nodes in a graph. They iteratively assign labels to nodes until they reach a stable state, which allows community detection. Random walk algorithms use random traversal of the graph, taking into account the probability of moving to neighboring nodes. Nodes that frequently visit each other are assigned to the same community.

The spectral method uses spectral analysis of a graph's adjacency matrix or Laplace matrix to find the directions with the most significant variability. It then clusters the nodes based on these directions, which leads to community detection. Choosing an appropriate method depends on many factors, such as the graph's structure, data availability, and the research goal. Many advanced variants and hybrid approaches also integrate different techniques to obtain more precise results.

The Label Propagation method is a simple and effective algorithm for detecting communities in graphs. It involves iteratively assigning labels to nodes in the graph based on the labels of their neighbors. In each iteration, a node adopts the dominant label among its neighbors. This iterative process continues until the assigned labels are no longer changed.

Figure 10. *Communities in the graph obtained using the Label Propagation method*



Source: *Own creation.*

Research on graph communities has used similarity matrices and tested various cutoff parameters to identify emerging groups. The matrices used include the cosine similarity matrix, and the similarity matrix was calculated using the proposed methodology based on the distribution function. The research was conducted for two different cut-off parameters: 0.5 and 0.75. Example analysis results are presented in Figure 10.

3. Conclusions

The presented methodology for analyzing customer data and creating a recommendation system based on cosine similarity between them draws several important conclusions.

First, using cosine similarity as a measure of closeness between customers is an effective technique for identifying similar groups of customers. This method makes it possible to detect customers with similar preferences, shopping habits, or demographic characteristics, which can be used to personalize the offer and marketing activities.

Secondly, an essential process element is appropriate data preparation, including combining customer features into one representation and calculating a cosine similarity matrix. Proper data analysis and processing are necessary for accurate results and practical recommendations.

It is also worth noting that creating a recommendation system requires a proper understanding of customer needs and expectations. Therefore, it is important to constantly monitor and evaluate the system's effectiveness and adapt it to changing market conditions and customer preferences.

The work's results emphasize the importance of using data analysis methods and artificial intelligence in marketing and customer relationship management. The developed tools and techniques can help the company better understand its customers and more effectively adapt its offer to their needs, which contributes to increasing customer loyalty and improving business results.

References:

Chen, L., Zhang, H. 2022. Data-Driven Customer Profiling and Recommendation in Retail Industry. IEEE Transactions on Big Data, 9(4), 567-580.

Gupta, S., Kumar, A. 2024. A Novel Approach to Customer Segmentation and Targeted Marketing Using Graph-Based Recommendation Systems. Expert Systems with Applications, 185, 112345.

Johnson, R., Patel, S. 2024. Enhancing Customer Relationship Management through AI-Powered Recommender Systems. International Journal of Data Science and Customer Analytics, 7(1), 45-60.

Kulisz, M., Kłosowski, G., Rymarczyk, T., Słoniec, J., Gauda, K., Cwynar, W. 2024. Optimizing the Neural Network Loss Function in Electrical Tomography to Increase Energy Efficiency in Industrial Reactors. Energies, 17(3), 681.

Rymarczyk, T., Sikora, J. 2024. Some More on Logarithmic Singularity Integration in Boundary Element Method. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, 14(1), 5-10.

Smith, J., Brown, A. 2023. Effective Customer Segmentation Using Cosine Similarity-Based Recommender Systems. Journal of Artificial Intelligence in Marketing, 15(2), 87-102.

Thalassinos, E.I., Venediktova, B., Staneva-Petkova, D., Zampeta, V. 2013. Way of banking development abroad : branches or subsidiaries. International Journal of Economics & Business Administration, 1(3), 69-78.

Wang, Y., Li, X. 2023. Utilizing Machine Learning Techniques for Personalized Marketing: A Case Study in E-commerce. Journal of Marketing Analytics, 12(3), 210-225.

Zampeta, V., Chondrokoukis, G. 2023. A Comprehensive Approach through Robust Regression and Gaussian/Mixed-Markov Graphical Models on the Example of Maritime Transportation Accidents: Evidence from a Listed-in-NYSE Shipping Company. Journal of Risk and Financial Management, 16(3), 183.