Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 3 (2023)

Customer Segmentation using Machine Learning Techniques

R. Amutha[1], Aazmaan Ahmed Khan[2]

[1]*Associate Professor, Department of Information Science and Engineering*

AMC Engineering College, Bangalore-560083, Karnataka, India

[2]*PG Student, Department of Information Science and Engineering*

AMC Engineering College, Bangalore-560083, Karnataka, India

Abstract

The objective of this customer segmentation endeavor is to utilize machine learning methodologies, Python programming, and the Streamlit framework in order to deliver individualized suggestions pertaining to savings plans, loans, and strategies for wealth management. Through the division of customers according to their demographic data, purchase patterns, and online navigation tendencies, this initiative empowers financial establishments to gain deeper insights into their customer demographics and customize their services to cater to specific requirements.

The project involves several stages, including data collection, data preprocessing, feature selection, model training, and evaluation. The acquired customer data is refined to resolve any inconsistencies or missing information. Important characteristics are identified, and an exploratory analysis of the data is conducted to uncover hidden patterns and relationships. We then select an appropriate machine learning algorithm, such as K-means clustering, hierarchical clustering, or Gaussian Mixture Models, to categorize the customer base into distinct clusters.

After the completion of model training and assessment, a Streamlit-based web application is constructed, offering an interactive platform for users. This application permits users to input their particulars and obtain tailored suggestions grounded in their designated segment. The suggestions encompass fitting savings schemes, loan alternatives, and wealth management approaches that harmonize with their financial objectives and risk appetite.

The deployed application facilitates a seamless user experience, providing real-time recommendations and insights. Continuous improvement of the model and application is encouraged through user feedback, allowing for refinement and better customization of recommendations over time.

Keywords— *Customer Segmentation; Machine Learning Methodologies; python programming; Streamlit; K-Means*

Introduction

Customer segmentation is the practice of categorizing a customer base into separate groups or segments based on common traits, behaviors, or preferences. This process entails examining different data elements, including demographic data, purchase records, online activity, and psychographic characteristics, in order to detect trends and commonalities within the customer population.

In the competitive landscape of today's business world, understanding customers and efficiently reaching them with personalized marketing strategies is essential for a company's success. Customer segmentation, the practice of dividing a customer base into distinct groups based on common characteristics and behaviors, provides

organizations with a powerful tool to tailor their marketing efforts and enhance customer interaction. Advances in machine learning have made this approach increasingly precise, effective, and data-driven over time.

To effectively group customers for targeted marketing, it's vital to identify key differentiating factors that set them apart. When formulating customer segmentation strategies, aspects such as customer demographics, location, psychographics, and behavioral tendencies are taken into account.

Literature Survey

In the contemporary business environment, there is a growing need for an AI-driven customer profiling system. As competition intensifies and communication costs rise, and with the challenge of customer retention becoming more pronounced, businesses are redirecting their attention from acquiring new clients towards nurturing and maintaining their current customer pool. Building lasting customer relationships offers numerous financial advantages, including repeat purchases, reduced advertising expenses per customer, and organic customer base expansion through referrals. This project's objective is to create a consumer profiling system using machine learning techniques to improve automated marketing, boost sales, and enhance overall customer engagement. To achieve this goal, it is planned to solve the following research tasks:

- Data collection
- Study of machine learning techniques
- Define the format of the client profile, types, and attributes that describe them
- Analysis and formation of customer data
- Summarize and organize foreign experience in enhancing the profile of clients
- Perform an evaluation of existing methods for researching the profile of clients, and pinpoint the most efficient ones suitable for business

2.1 Customer Classification In the business landscape, competition has been on the rise as companies aim to satisfy the wants and needs of their current customer base while also expanding their clientele. Puwanenthiren[6] et al. elaborate that understanding and catering to individual customer preferences and objectives can be a complex undertaking. This complexity stems from the wide range of demands, preferences, demographics, sizes, and characteristics among various customers. Treating every consumer uniformly is no longer considered effective. To tackle this challenge, the practice of market or customer segmentation has been embraced, wherein customers are grouped into smaller segments based on shared market traits or behaviors. T. Nelson[9] et al. further explain that customer segmentation involves the process of categorizing the market into distinct groups.

2.2 Big Data Recently The realm of Big Data research has experienced a significant surge in recent times. The term 'big data' encompasses a broad range of structured and unstructured data that poses challenges to traditional analysis methods and algorithms. Enterprises amass extensive datasets related to their customers, suppliers, and operational activities. Millions of interconnected sensors are strategically deployed in real-world settings via devices like mobile phones and vehicles, capturing data relevant to sensing, manufacturing, and communication processes. McKinsey[10] et al. elaborate that this vast reservoir of information has the potential to enhance predictive capabilities, drive cost efficiencies, improve operational effectiveness, and drive advancements in various domains, including traffic management, weather forecasting, disaster preparedness, finance, fraud prevention, business transactions, national security, education, and healthcare. The characteristics of big data are often summarized by the three Vs: volume, variability, and velocity, with the addition of veracity and value, resulting in a comprehensive 5Vs framework

2.3 Data repository Data collection involves gathering and measuring information with the aim of assessing specific changes within an established system, enabling the investigation of pertinent questions and the evaluation of outcomes. A.K. Jain[12] et al. elucidate that data collection is an integral component of research across various domains, encompassing the physical and social sciences, humanities, and business. The

overarching objective of data collection is to acquire high-quality evidence that guides the analysis in generating well-founded and informative responses to the posed inquiries. In this study, the data was sourced from the UCI machine learning repository.

2.4 Clustering Data clustering entails the grouping of data points within a dataset based on shared characteristics. Various algorithms exist for this purpose, each suited for specific conditions. As noted by Sulekha Goyat[7] et al., there is no one-size-fits-all clustering algorithm, emphasizing the importance of selecting the most suitable clustering methods. In this study, three different clustering algorithms were employed, leveraging the Python scalar library.

2.5 K-mean The K-means algorithm is widely acknowledged as a prominent classification technique. This clustering approach revolves around centroids, where each data point is assigned to a predefined cluster within the K-algorithm framework. The clustering of data points reveals inherent patterns that provide valuable insights for informed decision-making. There are various strategies for implementing K-means, and for this study, we will employ the elbow method.

2.6 Spectral Clustering Spectral clustering stands as a robust method applicable for customer segmentation, relying on similarities or connections among customers. This technique harnesses graph theory and spectral analysis. Employing eigenvalues and eigenvectors of a similarity matrix derived from customer data, spectral clustering forms a graph-based clustering algorithm.

2.7 Gaussian mixture model GMM serves as a probabilistic model positioning the data distribution within each segment adheres to a Gaussian (normal) distribution. GMM-driven clustering presents numerous benefits for customer segmentation. It establishes a probabilistic structure enabling soft assignments, accommodating instances potentially belonging to multiple clusters. GMM adeptly captures intricate data distributions, handles overlapping clusters, and proves suitable for datasets encompassing continuous features.

2.8 DBSCAN DBSCAN (Density-Based Spatial Clustering of Applications with Noise) represents a density-centric clustering technique viable for customer segmentation. Notably adept at identifying clusters of diverse shapes and managing noisy data, DBSCAN excels in accommodating outliers. Unlike alternative clustering methods necessitating the predefinition of cluster numbers, DBSCAN autonomously deduces the appropriate number of clusters via data density assessment.

2.9 Agglomerative Clustering Agglomerative clustering stands as a hierarchical method suitable for customer segmentation. Employing a bottom-up strategy, it initiates with individual data points and gradually combines them into clusters according to their similarities. This technique presents a flexible and comprehensible approach for customer segmentation. It furnishes a hierarchical arrangement that facilitates the comprehension of nested clusters across different levels. Agglomerative clustering demonstrates adaptability, accommodating diverse data types and distance metrics. Additionally, it yields insights into inter-cluster relationships, enabling enhanced comprehension of customer segments.

This study is built upon extensive research conducted by both domestic and international scholars in fields related to the market economy, management, marketing, consumer behavior, and brand loyalty management. As described by C. Zopoundis et al.[21], the investigation employed a variety of methodologies, including marketing, economic, and statistical analyses, encompassing both quantitative and qualitative approaches. The study adhered to the principles of consistency and progression. Expert methodologies were utilized to validate critical aspects of the dissertation, following the framework articulated by authors H. Muller and U. Hamm. The initial stages encompass segmentation, marketing, and customer data analysis, with the collected data suitably prepared for profiling and analysis. H. Muller et al.[22] clarify that the novelty of this research lies in the development of scientific and methodological principles and recommendations directed towards the establishment and implementation of an AI-based customer profiling framework. Moreover, the study identifies the most effective techniques for studying customer profiles and enhancing customer loyalty within Kazakhstani

enterprises. The findings of this research will provide valuable insights for companies seeking to enhance their relationship marketing strategies and elevate sales performance through data-driven customer profiling.

In the field of customer segmentation, researchers have been investigating different algorithms for dividing customer data. Much of the research has been concentrated on analyzing customer purchase histories and buying behaviors to identify segments. T. Jiang et al.[23] emphasize the importance of integrating customer segmentation and buyer targeting to enhance marketing effectiveness. These activities are integrated into a systematic approach, but challenges arise in optimizing them cohesively. To address this issue, the authors propose the K-Classifiers Segmentation algorithm, which prioritizes resource allocation to customers with the potential for higher returns.

Numerous researchers have explored a wide array of methods for segmenting customers in their studies. Furthermore, the authors introduce a direct clustering approach for grouping customers, departing from computed statistics to include transactional data from multiple customers. Recognizing the computational complexity of achieving an optimal segmentation solution, a problem known as NP-hard, Tuzhilin presents alternative sub-optimal clustering methods. The study then proceeds to empirically evaluate the customer segments obtained through direct grouping, demonstrating their superior performance compared to statistical approaches.

KR Kashwan et al.[24] introduced a K-means algorithm integrated with a statistical tool, presenting a model designed for continuous analysis and an online framework tailored to an e-commerce entity for sales prediction. This computer-based intelligent system uses a clustering approach for market segmentation, enabling swift decision-making by promptly delivering results to managers. PQ Brito emphasized the importance of customized advertising and manufacturing strategies across diverse industries, given the complexity of identifying precise customer preference patterns within extensive product ranges. Accordingly, they advocated the use of two data mining methods, clustering, and sub-cluster discovery, to enhance customer segmentation and deepen the understanding of preferences. X He and C Li proposed a comprehensive strategy centered on enhancing customer lifetime value (CLV), satisfaction, and behavior analysis. Their research underscores the role of segmentation in identifying diverse customer needs and expectations, ultimately leading to improved service delivery. A Sheshasaayee developed an innovative integrated segmentation approach by combining RFM (Recency, Frequency, Monetary) and LTV (Life Time Value) methods. This two-phase strategy begins with a statistical method in the first phase, followed by clustering in the subsequent phase. The primary objective is to implement K-means clustering within the two-phase model, complemented by the utilization of neural networks to refine segmentation outcomes.

MT Ballestar et al.[25] delve into the significance of customer engagement in cashback utilization and assess customer behavior on a social network platform. They introduce a model that incorporates social network analysis to shed light on marketing aspects such as loyalty, communication, customer development, and engagement, underscoring the interdependence of customer positions within an organization. W Qadadeh suggests evaluating data analysis algorithms, including K-means for clustering and Self-Organized Maps for more detailed clustering with visualization. They advocate for the integration of diverse segmentation techniques guided by experts to drive organizational advancement, particularly in fields like insurance. This approach also involves a deep analysis of segment attributes and customer behavior within any customer relationship management dataset. AJ Christy underscores the pivotal role of segmentation in understanding customer needs and identifying potential customers for organizational satisfaction. They employ segmentation through RFM analysis and extend their approach to incorporate other algorithms like K-means and RM K-means, resulting in subtle adjustments to K-means clustering.

Hwang, Y. et al.[4], describes the Elbow Method as a technique employed to determine the number of centroids (k) in a k-means clustering algorithm. In this method, the k-value is obtained by iterating continuously from k=1 to k=n (where n is the hyperparameter chosen according to specific requirements). For each k-value, the within-cluster sum of squares (WCSS) is calculated.

Problem Statement

This study on customer segmentation seeks to evaluate the usage patterns of approximately 9,000 active credit card users during the preceding six months through the application of machine learning methods. The primary aim is to classify these customers according to their behavioral patterns and preferences, culminating in the development of a Streamlit application designed to provide personalized suggestions pertaining to loan offerings, wealth management strategies, and savings approaches.

The current challenge involves comprehending and categorizing the diverse array of active credit card users based on their usage patterns. Our aim is to identify discrete customer segments characterized by unique attributes and preferences, achieved by analyzing factors such as transaction records, spending tendencies, repayment conduct, and other relevant data.

Accurately segmenting clients demands intricate processing and analysis of the extensive credit card usage dataset, extracting valuable attributes, and applying machine learning algorithms. For comprehensive and personalized recommendations, the segmentation process must encompass not only transactional behaviors but also additional factors such as demographics, income levels, and credit histories.

The user-friendly interface of the Streamlit application developed within this project will serve as the conduit for interacting with the segmentation model and generating tailored recommendations. These recommendations will be curated to align with the financial goals, risk tolerance, and distinct needs of each customer segment. This encompasses guidance on appropriate lending options, wealth management strategies, and personalized plans for savings and investments.

Financial institutions can offer tailored financial goods and services by solving this issue and gaining important information into their consumer base. By providing individualized recommendations that are catered to each customer's specific needs and interests, this segmentation method improves customer happiness.

1. PROPOSED METHODOLOGY

The system facilitates the training of the dataset for consumer segmentation. The proposed approach for customer segmentation aims to effectively classify customers into distinct segments through the application of machine learning techniques, considering their attributes, behaviors, and preferences. Our solution empowers businesses to employ advanced algorithms for the execution of marketing strategies that enhance both customer satisfaction and overall company performance. Furthermore, it furnishes valuable insights into the consumer base.

4.1 Data Collection and Integration:

- Gather relevant customer information gathered from diverse origins, including demographics, transaction history, website interactions, social media data, and customer feedback.
- Integrate and preprocess the data, ensuring data quality, consistency, and privacy.
- Perform feature engineering to extract meaningful and relevant features from the collected data.
- Machine Learning Model Selection:
- Determine appropriate machine learning algorithms for customer segmentation, taking into account the data characteristics and the specific task.
- Investigate diverse algorithms such as clustering algorithms (such as k-means, DBSCAN, agglomerative clustering), dimensionality reduction techniques (like PCA, t-SNE), or ensemble approaches (like random forests and gradient boosting).

4.2 Training and Validation:

- Divide the preprocessed data into training, validation, and testing sets.
- Train the chosen machine learning models with the training data, fine tuning hyperparameters and

optimizing performance metrics.

- Validate the trained models using the validation set, ensuring their effectiveness and generalization capabilities.

4.3 Customer Segmentation:

- Apply the trained machine learning models to the testing set or new data to segment customers into distinct groups.
- Utilize clustering algorithms or other relevant techniques to identify similarities and patterns among customers.
- Categorize customers into fitting segments according to their feature representations and clustering results.

4.4 Interpretation and Analysis:

- Analyze the generated customer segments to understand their characteristics, preferences, and behaviors.
- Interpret the features and attributes that contribute to the differences and similarities between segments.
- Extract actionable insights from the segments, such as identifying high-value customers, understanding customer needs, or discovering untapped market opportunities.

4.5 Visualization and Reporting:

- Visualize the customer segments and their characteristics using charts, graphs, or other visualization techniques.
- Develop interactive dashboards or reports to present the segmentation results to stakeholders.
- Provide intuitive and user-friendly interfaces to enable stakeholders to explore and interact with the segmented customer data.

4.6 Deployment and Integration*:*

- Implement the developed customer segmentation system into a production environment, considering scalability, performance, and security requirements.
- Integrate the system with existing customer relationship management (CRM) systems, marketing automation platforms, or data analytics frameworks.

4.7 Continuous Improvement and Adaptation:

- Collect feedback from stakeholders and users to evaluate the effectiveness and importance of the customer segmentation system.
- Continuously monitor and update the system as new data becomes available or business requirements evolve.
- Stay informed about the latest advancements in machine learning and customer segmentation techniques to incorporate new methodologies or enhancements into the system.Authors and Affiliations.

System Architecture

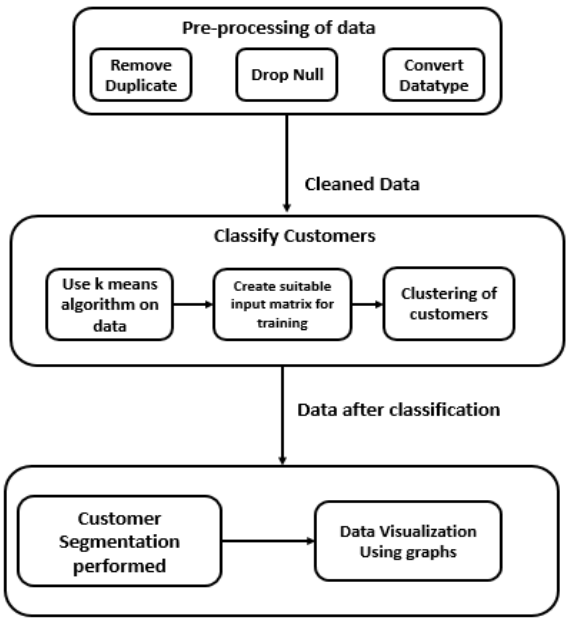

Fig 4.1 System Architecture for Customer Segmentation

2. Performance Analysis

5.1 Input

The provided dataset compiles the usage patterns of approximately 9,000 active cardholders over the past six months. This file is structured at the customer level and encompasses 18 behavioral variables.

"D:\Customer Segmentation\MarketSegmentation-main\Customer Data.csv"

Customer Segmentation is performed and provides recommendations like saving plans, loans, wealth management, etc. on target customers groups.

Input Values considered for segmentation

Balance – 40.900749

Balance Frequency – 0.818182

Purchases – 95.40

One-off Purchases – 0.00

Installment purchases – 95.40

Cash Advance – 0.000

Purchases Frequency – 0.166667

One-off Purchases Frequency – 0.00

Purchases Installments Frequency – 0.083333

Cash advance Frequency – 0.000

Cash Advance TRX - 0

Purchase TRX - 2

Credit Limit – 1000

Payments – 201.802084

Minimum Payments – 139.509787

PRC Full Payments – 0

Tenure - 12

5.2 Output

Finding the K value using the elbow method

The steps can be defined as follows:

1. Perform K-Means clustering for a range values of K, iterating from 1 to 10 clusters.
2. For each value of K, compute the total within-cluster sum of square (WCSS).
3. Create a plot showing the relationship between WCSS and the no of clusters K.
4. The presence of an inflection point (knee) in the plot is generally considered as an indication of the suitable number of clusters.

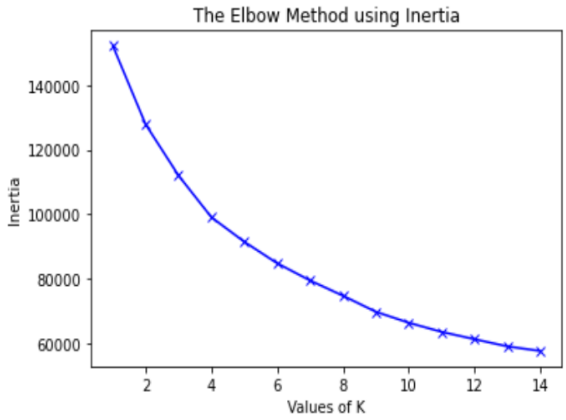

Fig 5.1 Finding k value by Elbow Method

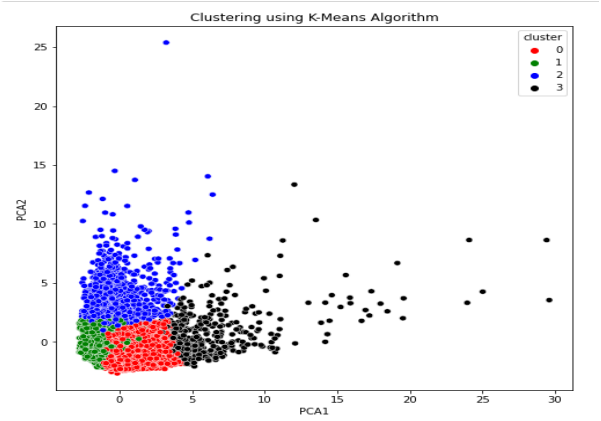

Fig 5.2 Visualizing the clustered Data Frame

3. CONCLUSION

As a result, this project involving machine learning, Python, and Streamlit has successfully showcased its ability to deliver personalized suggestions across domains such as savings strategies, loans, and asset management. Harnessing the potential of data analysis and machine learning algorithms, financial establishments can attain valuable insights into their clientele, thereby customizing their offerings to address specific needs.

Segmentation procedure separated customers into various groups according on their demographic information, past purchases, and browsing patterns. By adopting this segmentation, financial institutions can better understand the many profiles and preferences within their customer base and provide individualized guidance and specialized solutions.

By employing this client segmentation strategy and providing relevant, personalised advice, financial businesses may boost customer satisfaction. Customers benefit from receiving specialized advice on asset management strategies, financing options, and savings plans that fit their financial objectives and risk tolerance. This personalized approach will help customers become more loyal and trustworthy, which will ultimately increase engagement and improve the financial institution's bottom line.

Further, this project provides a framework for future expansion and development. The accuracy and applicability of the recommendations can be improved even more by adding extra attributes and data sources.

The project's extension to include more financial services and products makes it possible to take a thorough and all-encompassing approach to customer-centric financial management.

References

[1] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited.

[2] Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.

[3] Hong, T., Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. Expert System Applications, 39 (2), 2127-2131.

[4] Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using python and r. S.l: Packt printing is limited.

[5] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[6] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[7] Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011.

[8] By Jerry W Thomas. 2007. Accessed at: www.decisionanalyst.com on July 12, 2015.

[9] T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. AnuManufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.

[10] McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015.

[11] Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdxwhitepaper-090413.pdf July 14, 2015.

[12] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3.

[13] D. Nurseitov, K. Bostanbekov, D. Kurmankhojayev, A. Alimova, A. Abdallah, R. Tolegenov, Handwritten kazakh and russian (hkr) database for text recognition, Multimedia Tools and Applications 80 (21) (2021) 33075–33097.

[14] N. Toiganbayeva, M. Kasem, G. Abdimanap, K. Bostanbekov, A. Abdallah, A. Alimova, D. Nurseitov, Kohtd: Kazakh offline handwritten text dataset, Signal Processing: Image Communication 108 (2022) 116827.

[15] A. Abdallah, M. Hamada, D. Nurseitov, Attention-based fully gated cnn-bgru for russian handwritten text, Journal of Imaging 6 (12) (2020)

[16] G. A. Daniyar Nurseitov, Kairat Bostanbekov, Maksat Kanatov, Anel Alimova, Abdelrahman Abdallah, Classification of Handwritten Names of Cities and Handwritten Text Recognition using

Various Deep Learning Models, Advances in Science, Technology and Engineering Systems Journal 5 (5) (2020) 934–943. doi:10.25046/aj0505114.

[17] M. Mahmoud, M. Kasem, A. Abdallah, H. S. Kang, Ae-lstm: Autoencoder with lstm-based intrusion detection in iot, in: 2022 International Telecommunications Conference (ITC-Egypt), IEEE, 2022, pp. 1–6.

[18] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, F. Sabrina, Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset, IEEE Access 9 (2021) 140136–140146.

[19] B. Waschneck, A. Reichstaller, L. Belzner, T. Altenm¨uller, T. Bauernhansl, A. Knapp, A. Kyek, Optimization of global production scheduling with deep reinforcement learning, Procedia Cirp 72 (2018) 1264–1269.

[20] M. A. Hamada, A. Abdallah, M. Kasem, M. Abokhalil, Neural network estimation model to optimize timing and schedule of software projects, in: 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), IEEE, 2021, pp. 1–7.

[21] C. Zopounidis, New trends in banking management, Springer Science & Business Media, 2002.

[22] H. M¨uller, U. Hamm, Stability of market segmentation with cluster analysis–a methodological approach, Food Quality and Preference 34 (2014) 70–78.

[23] T. Jiang, A. Tuzhilin, Improving personalization solutions through optimal segmentation of customer bases, IEEE transactions on knowledge and data engineering 21 (3) (2008) 305–320.

[24] K. R. Kashwan, C. Velu, Customer segmentation using clustering and data mining techniques, International Journal of Computer Theory and Engineering 5 (6) (2013) 856.

[25] M. T. Ballestar, P. Grau-Carles, J. Sainz, Customer segmentation in e-commerce: Applications to the cashback business model, Journal of Business Research 88 (2018) 407–414.