2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)
Qazvin Islamic Azad University, Feb. 29, 2024

Data Mining and Cloud Computing for Customer Pattern Analysis and Value Maximization

1st Mohammad Sadegh Sirjani
Department of Computer
Ferdowsi University of Mashhad
Mashhad, Iran
mohammadsadegh.sirjani@mail.um.ac.ir

2nd Seyed Amir Mousavi
Department of Computer
Ferdowsi University of Mashhad
Mashhad, Iran
amirmousavi@mail.um.ac.ir

3rd Mostafa Sadeghi
Faculty of Computer Engineering,
Najafabad Branch, Islamic Azad
University, Najafabad, Iran.
Mostafa13h@gmail.com

2024 10th International Conference on Artificial Intelligence and Robotics (QICAR) | 979-8-3503-4887-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/QICAR61538.2024.10496623

***Abstract*—In the contemporary business landscape, companies are seeking efficient methods to analyze customer behavior and extract actionable insights to foster customer relationships and drive business growth. This paper presents a novel approach that combines the strengths of cloud computing, software engineering, and data mining techniques to analyze customer patterns and optimize the value of the customer life cycle. By employing data mining techniques, including clustering, classification, association analysis, and predictive modeling, businesses can identify customer patterns and behaviors. Through the extensive analysis of customer data, organizations can uncover concealed patterns, preferences, and trends, which facilitate informed decisions regarding customer acquisition, retention, upselling, and personalized marketing strategies, ultimately maximizing the value of the customer life cycle. The integration of cloud computing, software engineering, and data mining enables businesses to leverage advanced analytics and extract valuable insights from customer data. This approach enables organizations to gain a comprehensive understanding of customer patterns and behaviors, thereby facilitating targeted marketing campaigns, personalized customer experiences, and improved customer satisfaction.**

***Keywords*—Data mining, Software engineering, Customer pattern analysis, Cloud computing, optimization**

I. Introduction

Cloud computing and software engineering, as critical technologies in the digital era, are highly effective instruments for data analysis and the detection of customer tendencies. By combining these two technologies with data mining, there arise exceptional prospects to identify patterns, trends, and customer conduct. [1].

Data mining is a method of analytical analysis which employs a range of algorithms and techniques to uncover concealed information and establish intricate connections within a dataset. [2]. Utilizing cloud computing, we are able to handle a vast quantity of information at a rapid pace and with high levels of efficiency, enabling us to analyze large data sets. Furthermore, the software engineering and the underlying architecture of the technology boast advanced tools and systems for data extraction, transformation, and analysis. [3].

Through the process of data mining and analyzing customer patterns, it is possible to gain a more accurate understanding of the purchasing behaviors, preferences, and needs of customers. This information can then be utilized by companies to develop effective strategies for attracting new customers, retaining existing ones, and increasing overall customer satisfaction. [4].

Utilizing cloud computing affords the advantage of access to high processing power and scalability. This allows for the efficient and speedy processing of large volumes of data, as well as the ability to execute tasks in parallel. In addition, the implementation of effective systems for data storage and management, along with the application of software engineering and the architecture of the cloud, facilitates the execution of data mining algorithms. [5].

In a nutshell, the synergy between cloud computing, software engineering, and the spatial architecture with data mining opens up powerful opportunities for uncovering customer behavior and optimizing marketing and customer service approaches. This collaboration enables organizations to extract more value from their data and make more astute decisions based on customer trends and patterns. [6].

The synergy between cloud computing, software engineering, and architecture has given rise to innovative ideas in the realm of data mining. By leveraging cloud computing, software engineering, and the architectural design of the said space, sophisticated programs can be developed to analyze customer trends. Through the deployment of cloud computing, a substantial amount of customer data can be processed swiftly and with high capacity, and by utilizing the software engineering and architectural design of that area, advanced data mining algorithms and predictive models can be executed. [4-6].

Through the analysis of customer data via cloud computing, it is possible to discern purchasing patterns, preferences, and habits. By employing software engineering and the architecture of that domain, predictive models may be fashioned to forecast the future behavior of customers. This information may be utilized for strategic determinations pertaining to product development, the improvement of customer service, pricing strategies, and advertising campaigns, among other purposes.

By utilizing data analytics and cloud computing, it is possible to enhance the efficiency of marketing efforts and drive better results for businesses. By analyzing customer data, companies can gain insights into the needs and preferences of their target audience, thereby allowing them to customize their services and products more effectively. This personalized approach can lead to increased customer satisfaction, better business results, and reduced resource waste. Furthermore, cloud computing offers high scalability, making it easier to process large volumes of data and produce timely results. [7]. Utilizing software engineering and the architectural design of the cloud computing environment, it is feasible to develop and deploy efficient and scalable programs for data processing and analysis. Specifically, the integration of cloud computing, software engineering, and spatial architecture in data mining affords businesses with formidable tools for customer pattern analysis and prediction, enabling them to make informed strategic decisions and reinforce relationships with their clientele. [8]. In this investigation, a two-objective optimization model based on customer lifetime value was introduced to identify target customers for maintenance programs and to determine the relevant costs (customer contact and offer) for each customer. The two objective functions are designed to maximize customer lifetime value by implementing a retention program, while minimizing the costs of the program. The model was tested using real data from an insurance organization in the country, and the analysis was conducted in three stages. In the first stage, the probability of customers defecting was obtained through data mining. In the second stage, the customer lifetime value was calculated, and in the final stage, the optimization model was solved using the LP metric and GAMS software.

II. Ease of Use

A. Background

The field of information technology and the Internet have become increasingly important in modern society, particularly with regards to issues such as information security, processing speed, access, resource allocation, cost savings, and server maintenance. In response to these concerns, the field of technology has recommended cloud computing as a solution. Cloud computing is a model that provides on-demand access to a shared pool of configurable computing resources, such as networks, servers, storage space, applications, and services, with minimal involvement or management required from the service provider. [9-10].

Corporate cloud computing offers customers the option to pay for IT services on a subscription basis. This means that instead of purchasing servers or software licenses, businesses can access these resources by subscribing to them as a service. The use of cloud computing can free up space in existing facilities without the need for costly hardware upgrades, employee re-training, or software license upgrades. Ultimately, businesses and consumers only pay for the services they use. [2]. Consequently, it is a cost-efficient means of utilizing assets, managing cash, and providing technological aid. Several initiatives have been undertaken to improve the efficiency of the data center [3, 4, 5, 6]. A formal approach was devised to investigate the time-related variations in the hybrid cloud while considering delay limitations. [7]. Cloud security, which is also known as the security of cloud computing, is a subfield of the broader domains of computer, network, and information security [9,10]. The notion of cloud computing is straightforward yet has far-reaching ramifications. Rather than relying on a locally-based computer to store data, servers located online will protect applications, data, and the underlying security architecture from distant access. This approach enables users to use any internet-connected device, including smartphones, high-quality PCs, or even airport internet kiosks, to upload images, information, and media to a selected server, regardless of its location [11].

Despite the widespread interest in cloud computing and the numerous descriptions of the field that have been proposed, a consensus on how to define cloud computing remains elusive. Cloud computing continues to be a topic of considerable discussion [1]:

Cloud computing is a model of delivering computer-related services over the Internet, where users can access these services through subscription-based arrangements.

The advent of new methods for presenting and delivering IT has led to the immediate internet-based provision of goods, services, and solutions.

We aim to provide our clients with a single, integrated platform that offers a range of business-critical functions, including accounting, sales automation, and customer care, all hosted on a secure server.

Telecommunications providers have traditionally supplied dedicated point-to-point lines, but their scope of service has since expanded to include virtual private networks (VPNs) that offer comparable quality and cost. The cloud, represented in this diagram, denotes the line of demarcation between the user's responsibilities and the provider's purview. With the advent of cloud computing, the cloud metaphor has been expanded to encompass not only data centers, but also computer servers. By investing in cutting-edge data centers, Amazon has made significant contributions to the advancement of cloud computing [1].

B. Research methodology

Cloud computing offers scalable and immediate access to computational resources, allowing for the efficient processing of significant amounts of customer data. Organizations may harness the cloud infrastructure to perform intricate data analysis tasks at a rapid pace, thereby enabling the extraction of relevant insights from substantial and diverse customer datasets [12].

Typically, three tiers exist in a cloud computing architecture [13].

Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are three models of cloud service delivery that have gained widespread popularity in recent years.

Figure 1 illustrates the general structure of a cloud computing service, but it should be noted that the composition of this system is comprised of five distinct elements.

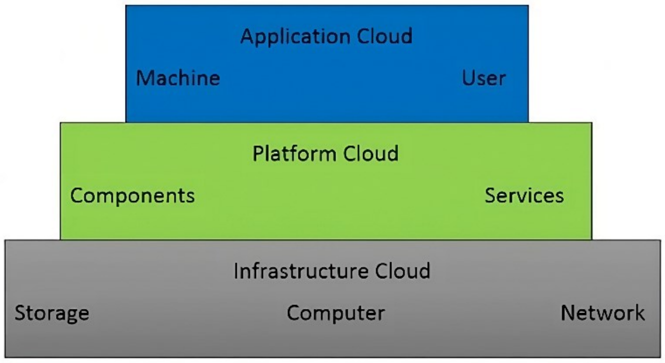

Figure1. cloud computing layers

Cloud computing refers to the use of computer software and/or hardware that relies on cloud computing for the delivery of applications, or that is specifically designed for the provision of cloud services. The technology is considered to be of little value in the absence of cloud computing. The characteristics that define this stratum are as follows:

The topic of discussion involves a comparison between software and hardware.

The cloud service recipient experiences ineffectiveness in the absence of cloud services.

The preceding list of devices includes a variety of operating systems, such as iOS, Android, and Windows Mobile, as well as lightweight alternatives like Zonbu, CherryPal, and GOS-based systems. Furthermore, there are numerous user and web browsers available, including Firefox, Google Chrome, and WebKit. [14].

The second level of cloud computing is Software as a Service (SaaS), which is characterized by the delivery of software applications over the internet.

Cloud application services, also known as software-as-a-service (SaaS), provide software delivery via the internet, thus eliminating the need to install programs on client devices. This method simplifies the maintenance and support processes. The fundamental attributes of such services are as follows:

Access to and management of commercial software via a network are the topic of discussion. It is important to maintain a formal tone while addressing this matter. [10]. The centralized management of activities is conducted at designated centers that are located separately from the customers' sites. This arrangement enables remote access to applications via web-based interfaces, and the software delivery model more closely resembles a one-to-many configuration, in which a single running version of the application is serving multiple tenants, rather than a one-to-one relationship [11].

A formal tone would be appropriate for the following statement: The centralized management of software updates and upgrades removes the need for downloading patches or upgrading. The application layer encompasses a range of software and services, such as P2P programs like Skype, web applications like Facebook and YouTube, security services like MessageLabs, software services like IBM Lotus and Google Labs, add-on services like Microsoft Online Services, and storage services. These companies were established with a focus on providing software services, and charge fees for user registration with the software installed on central servers. Users can access applications through the Internet. [14].

The third level of cloud computing architecture is denoted as Platform as a Service (PaaS).

Our platform provides a Platform-as-a-Service (PaaS) offering, which is a computing platform that is hosted on a cloud infrastructure and utilized as a service for cloud-based applications. This model operates as a service rather than a platform software [14].

Platform as a Service (PaaS) affords the opportunity for software expansion without requiring substantial investments in hardware and software acquisition and management. Furthermore, PaaS offers web hosting capabilities, thereby eliminating the need for software developers to bear the costs of developing new applications or improving existing ones [1].

The aforementioned service provides a software solution that enables the generation of services at a higher level. The platform encompasses a range of components, including middleware, integration capabilities, message exchange, information sharing, and connection establishment.

The fourth level of cloud computing, typically designated as Infrastructure as a Service (IaaS), is characterized by the provision of virtualized computing resources, including servers, storage, and networking, over the internet.

Infrastructure as a Service (IaaS) is a cloud computing model that provides a virtualized computer infrastructure as a service. This infrastructure is typically offered in the form of a platform and is procured as an outsourced service, rather than through the acquisition of hardware, software, data center space, or network equipment. The cost of the service is typically determined by the level of resource utilization and is therefore proportional to the degree of engagement with the public computing paradigm. This model represents a departure from the conventional virtual private server provision model, which is typically characterized by a virtualized computing environment. [6]

Broadly defined, Infrastructure as a Service (IaaS) encompasses the provision of virtualized computer infrastructure and platforms as a service. An example of such a service is Amazon Web Services (AWS). [8].

The fifth stratum of the network structure was the server echelon.

The term cloud computing infrastructure encompasses the physical components of cloud computing, including specialized hardware and software utilized to provide cloud services. Examples of such infrastructure include multicore processors and operating systems designed specifically for cloud computing [5].

The overall cloud computing architecture is composed of five discrete layers, and it is expected that individuals will conduct additional research into the evolution and progress of the initial and subsequent tiers, recognizing the importance of the remaining tiers for both service providers and developers during this transitional phase.

K nearest neighbor (KNN)

This algorithm classifies sample data by identifying the K most similar data samples in the learning set. It utilizes a dataset and an observation target, with each observation in the dataset featuring a set of variables and the target observation possessing a value for each variable. The determination of KNN is typically based on the Euclidean distance between the observation and the observation target [2-3].

III. Model evaluation

Once the predictive model had been developed, it was utilized to forecast the future conduct of customers. Ensuring the accuracy of the prediction model was a critical step in this process.

A. Lift and Gain diagram

The lift measure is a widely recognized method for evaluating the performance of classification models. It quantifies the difference in class distribution between the population and the model's selected group. The lift ratio is calculated as the change in class density, indicating the extent to which the model successfully identifies the target class. [6-7-8].

The lift function is directly proportional to the sample size. To calculate the lift, the software first sorted the records based on their likelihood of being classified as positive. The resulting lift values were plotted against various percentages of the sample size. The value of this rate decreased from left to right and reached a value of one when it was applied to the entire sample size. A higher rate indicates a more accurate classification model. It is expected that the density rate of the top 10% of society is higher than that of society as a whole.

Lift and Gain charts are graphical validation methods utilized for estimating and comparing the efficiencies of classification models. The Lift chart is often displayed in a cumulative format, known as the Gain chart form. The diagonal line drawn in the Gain diagram represents the horizontal axis of Lift=1 in the Lift diagram. Typically, Gain graphs gradually diverge from this natural line and follow a curved graph above the main diameter until they intersect with it again. Lift and Gain charts are commonly utilized to compare the performance and efficiency of various models.

$$lift = \frac{P(class_t \mid sample)}{P(class_t \mid population)} \qquad (1)$$

The correlation matrix was utilized to assess the performance of the classification models with binary target variables. The correlation matrix is an essential tool in this process.

Based on this matrix, four key parameters are defined: TP, FP, FN, and TN.

TP represents the number of positive elements that were correctly classified as positive.

FP represents the number of negative elements that were incorrectly classified as positive.

FN represents the number of positive elements that were incorrectly classified as negative.

TN represents the number of negative elements that were correctly classified as negative.

Based on the adaptation matrix, three critical metrics, namely sensitivity, detection, and accuracy, are delineated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \qquad (2)$$

$$Specifity = \frac{TN}{TN + FP} \qquad (3)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (4)$$

The ROC chart was utilized to depict the relationship between the false-positive and false-negative rates. The false positive rate is plotted on the horizontal axis, while the false negative rate is represented on the vertical axis of the graph.

An ROC plot is often utilized to assess the performance of a classifier, and the area under the plot (AUC) is typically evaluated. This metric provides a measure of the classifier's behavior, independent of the chosen threshold or cost of misclassification.

Above the baseline, the ROC plot indicated a model with commendable performance. The baseline, which denotes a 50% value, serves as a reference for evaluating models that perform worse than the random selection scenario.

In this study, the use of the overall error rate and lift diagram serves to validate the results, while the mathematical formula for Customer Lifetime Value (CLV) is expressed in English as follows:

CLV = (Average Annual Revenue per Customer) × (Customer Lifespan)

or

CLV = (Average Annual Net Profit per Customer) × (Customer Lifespan)

$$CLV = \sum_{t=0}^{T} \frac{m_t r_t}{(1+i)} \qquad (5)$$

The Average Annual Revenue or Net Profit per customer signifies the average yearly income or profit generated by each customer. The Customer Lifespan denotes the typical length of time that a client remains involved with the firm, which can be expressed in years or months.

Using this formula, it is possible to determine the CLV value for each customer, as well as make informed decisions regarding marketing strategies, service levels, customer retention, and customer acquisition. By calculating CLV, one can prioritize and focus on customers with high CLV, ultimately resulting in the extraction of greater value.

Building data mining models

Given the unbalanced nature of the data under review, the ratio of returning customers to non-returning customers is exceptionally low. Consequently, a method for balancing the data should be employed. In this regard, the placement method was utilized to sample the rare class (reverting), and the ratio of turning to the non-reverting class was adjusted to 1 to 2. As a result, a total of 5711 samples were collected, of which 1924 were related to returning customers.

In the present investigation, the CHAID decision tree, perceptron neural network, and K-nearest neighbor algorithms were employed. As depicted in Figure 2, the data mining process was implemented within the specified environment.

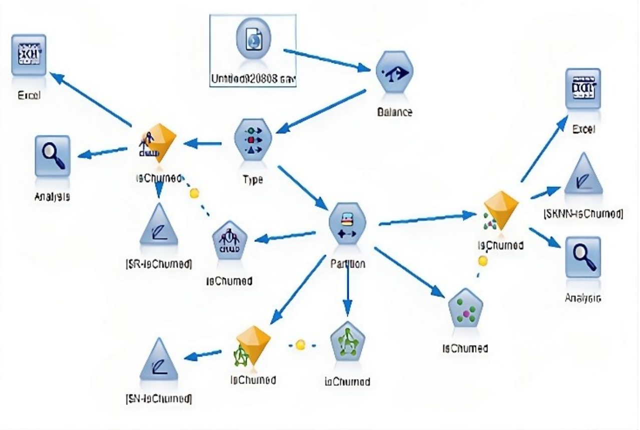

Figure 2. View of the data mining in a software environment

IV. VALIDATION

Figure 3 depicts the lift curve of the k-nearest neighbor method. The lift value is calculated by determining the ratio of the predicted parameter in various sections of the population (e.g., the first 20%) to the entire population. In this case, if the lift rate for the first 10% is greater than two, the classification model is deemed acceptable for the studied society. As shown in the figure, the lift value for the used model is higher than 2.5, indicating its validity in the category of customers.

$$CLV = \sum_{t=0}^{T} \frac{m_t r_t}{(1+i)} \qquad (6)$$

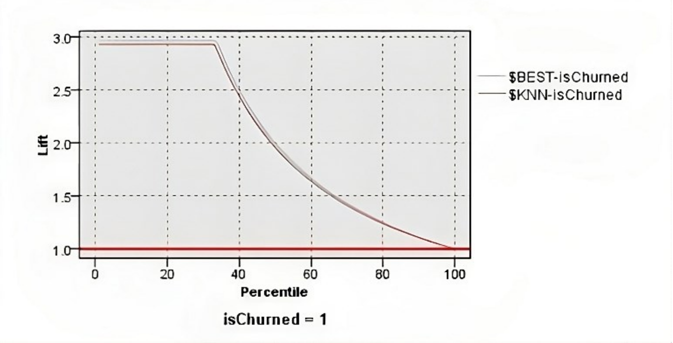

Figure 3 Lift diagram of the k-nearest neighbor method

A. *Calculating customer lifetime value*

To calculate the future customer lifetime value (CLV) for a given customer base, a linear regression model was employed. This model analyzed the customer's income data, serving as the response variable, and time as predictor variables. Subsequently, the r2 benchmark value was determined for each customer. The mean r2 was found to be 0.75. As shown in Figure 4, the distribution of CLV values for customers was found to be concentrated around a central tendency of 1820.00.

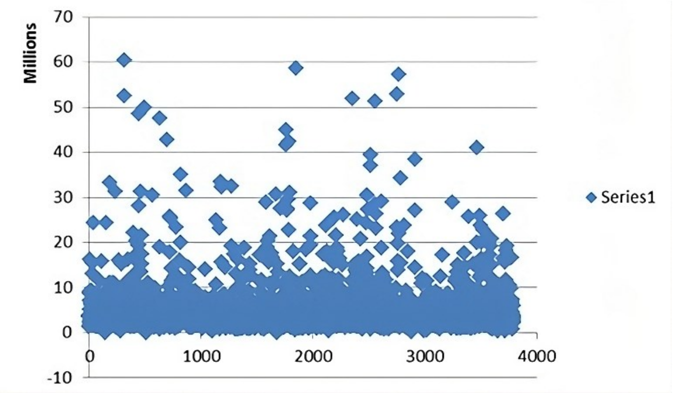

Figure 4 Distribution of customer lifetime value

The segmentation of the lifetime value of customers can be divided into four distinct groups, as depicted in Figure 5. The corresponding percentage of customers within each group is also indicated.

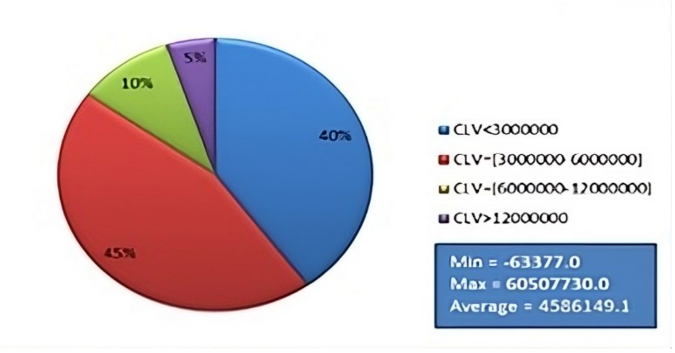

Figure 5 Lifetime value of customers

Table 1 displays the values obtained for the various criteria utilized in the tested algorithms. It is evident that the k-nearest neighbor method exhibited the best overall accuracy and lift criteria. A thorough examination of overall accuracy for different values of k was conducted, leading to the conclusion that k=1 provided the optimal solution. As a result, the outcomes for k=1 are regarded as the performance of the k-nearest neighbor method.

Table 1 Comparison of the results of classification methods

Lift 20%	Accuracy	Method
2.9	83.95%	decision tree
1.6	70.79%	neural network
2.5	97.33%	K nearest neighbor

The rates of sensitivity and detection parameters are as follows:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{1872}{1872 + 52} = 0.9 \qquad (7)$$

$$Specifty = \frac{TN}{TN + FP} = \frac{3762}{3762 + 24} = 0.9 \qquad (8)$$

The sensitivity parameter, expressed as a proportion, depicts the degree to which the positive elements of the population are accurately predicted in relation to their actual value. In this instance, the model has identified 97% of the defectors within the population.

Similarly, the specificity parameter represents the ratio of negative elements within society and highlights the percentage of genuine non-reverters that are correctly identified by the predictive model. The model has successfully identified 99% of non-reverting individuals.

CONCLUSION

Given the low cost of relocation within the industry, only short-term planning strategies are effective. This research represents the first instance of implementing a quantitative model with an optimal solution for this type of short-term planning.

In the field of turning in the industry, limited research has been conducted, with only one field's customer base being considered. No prior research has examined the behavior of customers across all contracts. This study has employed data mining techniques to determine the probability of customers switching due to multiple contracts.

Unlike other research that relies on the annual retention rate as a constant for all customers, this study calculates the customer lifetime value by considering the probability of each customer switching. Additionally, rather than using historical turnover rates to predict future turnover, this research has employed data mining tools to estimate the probability of customer churn in the next period.

REFERENCES

[1] Srinivas, M.A., Srinivas, M.K. and Varma, A.H.V., 2013. A study on cloud computing data mining. International Journal of Innovative Research in Computer and Communication Engineering, 1(5), pp.1232-1236.

[2] Fareed, M. and Al-Saedi, K.H., 2022, November. Proposal to enhance the information security system of the cloud using data mining techniques. In AIP Conference Proceedings (Vol. 2394, No. 1, p. 050012). AIP Publishing LLC.

[3] Mohbey, K.K. and Kumar, S., 2022. The impact of big data in predictive analytics towards technological development in cloud computing. International Journal of Engineering Systems Modelling and Simulation, 13(1), pp.61-75.

[4] Ali, M.H., Jaber, M.M., Abd, S.K., Alkhayyat, A. and Albaghdadi, M.F., 2022. Big data analysis and cloud computing for smart transportation system integration. Multimedia Tools and Applications, pp.1-18.

[5] Yu, L., Zheng, J., Shen, W.C., Wu, B., Wang, B., Qian, L. and Zhang, B.R., 2012, August. BC-PDM: data mining, social network analysis, and text mining system based on cloud computing. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1496-1499).

[6] Li, Q., 2022. Design of Customer Churn Early Warning System Based on Mobile Communication Technology Based on Data Mining. Journal of Electrical and Computer Engineering, 2022.

[7] He, Q. and He, H., 2020. A novel method to enhance sustainable systems security in cloud computing based on the combination of encryption and data mining. Sustainability, 13(1), p.101.

[8] Kawathekar, M., Niharika, Auti, R., Hudnurkar, M. and Sinha, M., 2022. Cloud Computing and Data Mining in E-Commerce. In CHANGING FACE OF E-COMMERCE IN ASIA (pp. 339-358).

[9] Dev, H., Sen, T., Basak, M. and Ali, M.E., 2012, November. An approach to protect the privacy of cloud data from data mining-based attacks. In 2012 SC Companion: High Performance Computing, Networking Storage and Analysis (pp. 1106-1115). IEEE.

[10] Lo'ai, A.T., Mehmood, R., Benkhlifa, E. and Song, H., 2016. Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access, 4, pp.6171-6180.

[11] Neaga, I. and Hao, Y., 2014. A holistic analysis of cloud-based big data mining. International Journal of Knowledge, Innovation and Entrepreneurship, 2(2), pp.56-64.

[12] Khare, A., 2014, March. Big data: Magnification beyond the relational database and data mining exigency of cloud computing. In 2014 Conference on IT in Business, Industry, and Government (CSIBIG) (pp. 1-6). IEEE.

[13] Chu, X., Cao, F., Jiao, L., Wang, J. and Jiao, Y., 2022. Optimal Allocation of Higher Education Resources Based on Data Mining and Cloud Computing. Wireless Communications and Mobile Computing, 2022.

[14] Wang, G. and Ding, Z., 2022. Application of the Data Mining Model in Smart Mobile Education. Mobile Information Systems, 2022.