Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning-Driven Customer Segmentation and Targeted Marketing: A Systematic Reviews

Lahcen Abidar[1*], Ikran EL Asri[2], Dounia Zaidouni[3], Abdeslam Ennouaary[4]
[1,2,3,4] *Mathematics and Computer Science, INPT, Madinat Al Irfane, RABAT, Morocco*
[1] *abidar.lahcen@inpt.ac.ma,* [2] *elasri@inpt.ac.ma,* [3] *zaidouni@inpt.ac.ma,* [4] *abdeslam@inpt.ac.ma*

ARTICLE INFO

Received: 17 Dec 2024

Revised: 19 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Customer segmentation is a cornerstone of marketing strategy, enabling businesses to gain valuable insights into the needs, preferences, and behaviors of their customers. Machine learning has emerged as a powerful tool for customer segmentation, leveraging advanced algorithms to group customers based on shared characteristics and behavioral patterns.

This study presents a systematic review of 82 research papers published over the past decade that utilize statistical, machine learning, and deep learning techniques for customer segmentation. We propose a comprehensive categorization methodology for machine learning-driven segmentation algorithms and address key challenges, including data quality and availability, algorithm selection, feature engineering, business context alignment, and ethical and legal considerations.

Our finding several that machine learning algorithms are extensively employed across various industries, particularly retail, banking, and tourism. While deep learning techniques demonstrate superior accuracy in segmentation performance, machine learning methods remain more commonly adopted due to their balance of effectiveness and practicality.

We also analyze and compare widely used techniques, summarizing key datasets and proposed models in a structured format. This research highlights the significant benefits of machine learning-driven customer segmentation for both businesses and customers, offering actionable insights for its effective implementation.

Keywords: Machine Learning Customer Segmentation Targeted Marketing Data-Driven Marketing Feature Engineering Consumer Behavior Algorithm Selection Marketing Optimization

INTRODUCTION

Customer segmentation is a pivotal aspect of marketing strategy, involving the division of diverse customer bases into smaller, more homogeneous groups based on shared characteristics [27]. Traditional segmentation methods—such as demographic, psychographic, and behavioral approaches—have long been employed to categorize customers. While effective to some extent, these conventional methods are limited by their inability to capture complex customer behaviors and their reliance on static, predefined segments.

Recent advancements in artificial intelligence (AI) and machine learning (ML) have profoundly transformed the landscape of customer segmentation, offering solutions that address the limitations of traditional approaches. Machine learning, with its ability to process vast datasets and uncover intricate patterns, has emerged as a game-changer, enabling businesses to analyze customer data with unprecedented accuracy and depth. This shift has revolutionized industries such as retail, finance, and tourism, facilitating more personalized and targeted marketing strategies.

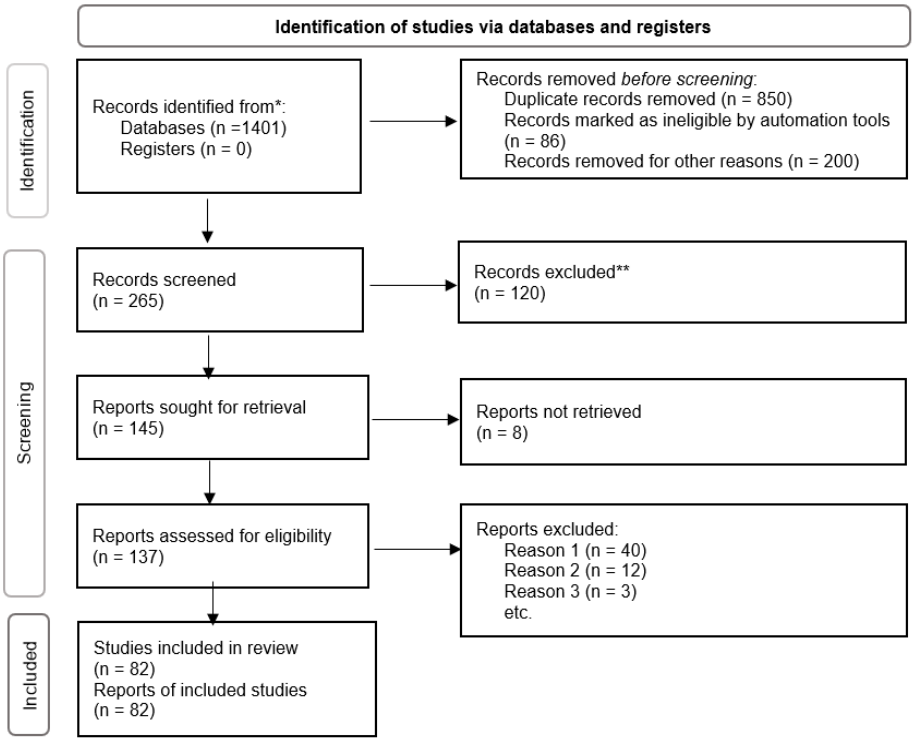
In this study, we present a comprehensive examination of the state-of-the-art in machine learning-driven customer segmentation and targeted marketing. First, we explore the machine learning algorithms most commonly employed in customer segmentation, including their capabilities and limitations. Next, we delve into practical applications

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

across various domains, highlighting how these techniques are implemented to achieve business objectives. Finally, we address the challenges and opportunities associated with the adoption of machine learning for customer segmentation, including issues of data quality, algorithm selection, and ethical considerations.

This work contributes to both academic and practical understanding by synthesizing recent advancements in machine learning-driven customer segmentation and providing actionable insights for implementation. It is particularly valuable for researchers, marketers, and business practitioners seeking to harness the potential of machine learning to enhance customer engagement and decision-making.



(a)

Figure 1: The PRISMA flow diagram [59]

The remainder of this paper is organized as follows: Section 2 outlines the study methodology. Section 3 introduces the key machine learning algorithms utilized for customer segmentation. Section 4 reviews their applications across industries. Section 7 discusses challenges and opportunities, while Section 8 concludes the study with recommendations and future research directions.

SURVEY METHODOLOGY

2.1. Methodology

We employed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reviewing approach in our paper. To begin our analysis, we employed five distinct search engines: Google Scholar, ACM, IEEEXplore, Springerlink, and ScienceDirect. While searching, we utilized the keywords "machine learning" or "deep learning" in conjunction with "customer segmentation" and "targeted marketing". We received 1401 items in all. Then, to move on, we utilized a filtering algorithm that considered the trade-off between publication year and citations. After deleting 850 duplicate records, 86 ineligible records, and 40 incomplete articles, we received 265 screened records. We eliminated 12 items because they were unrelated to the topic. Finally, we selected 82 studies for examination based on their relevance to the research issue, the precision of evaluation criteria, and the period since publication, and number of citations. The

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PRISMA flow diagram is shown in Figure 1.

2.2. Inclusion and exclusion criteria

In this work, we selected three inclusion criteria: (1) the relevance of the research topic, (2) the precision of assessment measures, and (3) the publication year and citations. Furthermore, articles that are duplicated, incomplete, published too early, unrelated to the topic, lack clear metrics, or have a low number of citations will be removed.

2.3. The datasets and techniques used in the reviewed papers

The primary datasets used in the papers under review are public and are obtained from a different public repository [5] [14]. Additionally, some studies have utilized their own data, such as (nilashi 2021) who used data from restaurants in Bangkok [76]. The machine learning approaches discussed in this review encompass various traditional models for customer segmentation, such as Support k-Nearest Neighbor (k-NN)[2] [14], Vector Machines (SVMs), Gradient Boost (XGBoost), Self-Organizing Maps, (SOM) Latent Class Analysis (LCA), Fuzzy Rule-Based Systems (FRBS), Decision Trees Support Vector Machines (SVM), Artificial Neural Networks (ANN), Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Additionally, neural network models, which fall under deep learning methods, are also discussed. These include (DNN) Deep Neural Networks, (RNN) Recurrent Neural Networks, (LSTM) Long Short-Term Memory, (CNN) Convolutional Neural Networks. Summary tables and bar charts providing an overview of the methods discussed in the reviewed papers are also provided.

COMPUTING APPROACHES

This section provides a brief overview of three major computing techniques used for customer segmentation and targeted marketing: machine learning, statistical, and deep learning (Figure 2).

3.1. Statistical Methods

Bayesian statistics is a branch of statistics that uses probability to make predictions or draw conclusions from data. The approach is based on the Bayes theorem, which describes how the probability of a hypothesis or model can be updated in light of new evidence [17]. The Logistic Regression (LR) approach is a statistical method for modeling the connection between a binary dependent variable and one or more independent variables. The purpose is to predict the probability that the dependent variable will be one of two different results [24]. LDA (Latent Dirichlet Allocation) is a probabilistic topic modeling technique used to discover latent topics in a large collection of documents. The approach assumes that each document is a mixture of topics, and each topic is a probability distribution over words [19]. HOSVD (Higher-Order Singular Value Decomposition) is a method for decomposing multi-dimensional arrays (tensors) into a core tensor and a set of orthogonal matrices. The approach is used in data compression, feature extraction, and data analysis [28]. These statistical methods have significantly contributed to the field of data analysis, enabling researchers and practitioners to extract meaningful insights and make informed decisions. Bayesian statistics provides a principled framework for incorporating prior knowledge and updating beliefs based on observed data, making it particularly valuable in decision-making under uncertainty [17]. LR has found widespread application in various domains such as healthcare, economics, and social sciences, where understanding and predicting binary outcomes are of utmost importance [24]. LDA, with its ability to uncover latent topics, has revolutionized text analysis and information retrieval, finding applications in fields such as natural language processing and social network analysis [19]. HOSVD, on the other hand, has proven instrumental in multidimensional data analysis, aiding in data compression, feature extraction, and pattern recognition tasks [28]. By leveraging these statistical methods, researchers and analysts can gain deeper insights into complex datasets, extract relevant information, and make data-driven decisions. Furthermore, the continuous advancements and applications of these methods contribute to the development of more sophisticated techniques and approaches in the field of data analysis. As data continues to grow in complexity and volume, these methods serve as valuable tools in extracting meaningful knowledge and enhancing our understanding of the world around us. Bayesian statistics, LR, LDA, and HOSVD offer powerful techniques for analyzing and interpreting data across various domains. These methods provide researchers with valuable tools to model relationships, discover latent patterns, and decompose

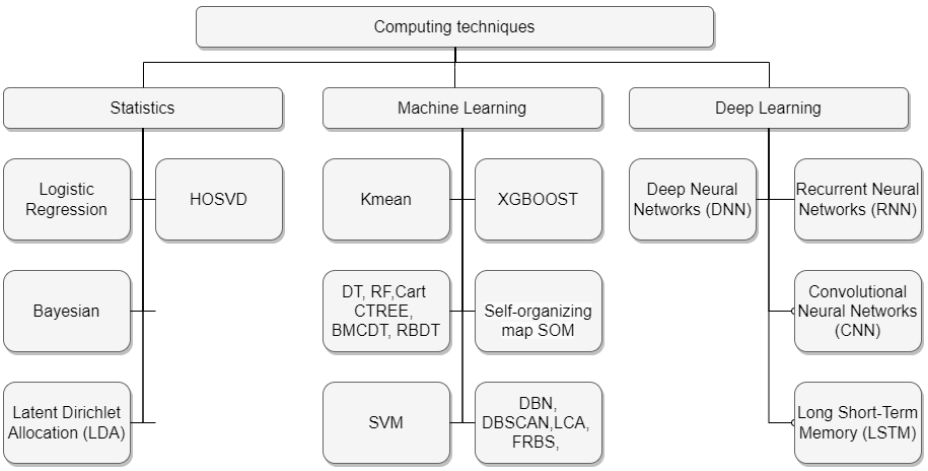
Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

complex datasets [17, 24, 19, 28]. As data analysis continues to evolve, these methods will remain essential in extracting insights and making informed decisions based on observed data.

3.2. Machine learning

XGBoost is an ensemble learning method that utilizes a set of decision trees to make predictions. It employs gradient boosting to train models and improve performance [21]. K-means is a clustering algorithm that partitions



(a)

Figure 2: Taxonomy Computing techniques

data into k clusters, where k is a pre-defined parameter. It is commonly used for unsupervised learning and data segmentation. Self-Organizing Maps (SOM) is a type of artificial neural network that uses unsupervised learning to cluster and visualize high-dimensional data in a lower-dimensional space [51]. Decision Trees (DT) are tree-like models used to make decisions. They consist of nodes representing features, branches representing decisions, and leaves representing outcomes. Various types of decision trees include regression trees [81], classification and regression trees (CART) [20], model-based trees (MOB), conditional inference trees (CTREE) [109], Bayesian model-based clustering decision trees (BMCDT) [80], and random binary decision trees (RBDT) [33]. Support Vector Machines (SVM, SVN, SVC) are supervised learning algorithms used for classification and regression analysis. They work by finding a hyperplane that separates the data into different classes [23]. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that groups closely packed data points together and identifies outliers as noise [34]. Latent Class Analysis (LCA) is a statistical model used to identify unobserved subgroups within a population based on observable variables. It finds applications in market research and social science [55]. Fuzzy Rule-Based Systems (FRBS) are rule-based systems that use fuzzy logic to make decisions. They are often used in control systems and decision-making tasks [107]. These machine learning and statistical techniques play significant roles in various domains, providing valuable tools for data analysis, pattern recognition, and decision-making processes. XGBoost, with its ensemble approach and gradient boosting, offers robust predictive modeling capabilities [21]. K-means clustering enables the identification of distinct clusters within data, aiding in unsupervised learning tasks and data segmentation. Self-Organizing Maps (SOM) leverage unsupervised learning to cluster and visualize complex high-dimensional data in a lower-dimensional space, facilitating pattern discovery and interpretation [51]. Decision Trees (DT) provide a flexible and interpretable framework for decision-making. They can be tailored to various tasks, such as regression, classification, and clustering, with different variants available [82, 20, 109, 80, 33]. Support Vector Machines (SVM, SVN, SVC) offer effective solutions for classification and regression analysis by finding optimal hyperplanes to separate data points [23]. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) excels in identifying dense regions and detecting outliers in datasets [34]. Latent Class Analysis (LCA) helps uncover hidden subgroups within populations based on observable variables, enabling deeper insights into complex datasets [55]. Fuzzy Rule-Based Systems (FRBS) with fuzzy logic provide a means to handle imprecise and uncertain information, making them valuable for decision-making tasks involving ambiguity

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[107]. These techniques, along with many others in the field of machine learning and statistics, contribute to the advancement of data analysis, modeling, and decision-making. By leveraging their strengths, researchers and practitioners can gain deeper insights, extract meaningful knowledge from data, and make informed decisions.

3.3. Deep Learning

Deep Neural Networks (DNN) are artificial neural networks (ANN) with numerous hidden layers. They are utilized for a variety of tasks such as image recognition, speech recognition, and natural language processing, [41]. Recurrent Neural Networks (RNN) are a form of artificial neural network that is used to analyze sequential input. RNNs use feedback loops to retain information from prior inputs. [87]. Long Short-Term Memory (LSTM) networks, a form of RNN, were created to address the problem of vanishing gradients. Speech recognition, natural language processing, and time series prediction are all common applications for LSTMs [42]. Convolutional Neural Networks (CNN) are neural networks that use a grid-like structure to process data, such as pictures. Convolutional layers are used by CNNs to extract features from input data [56]. These Deep learning approaches have transformed many fields by reaching cutting-edge performance in tasks such as picture classification, object identification, language translation, and sentiment analysis [3]. Their ability to learn hierarchical representations from data automatically has made them indispensable in many fields, enabling improvements in areas such as autonomous vehicles, healthcare, and recommendation systems.

CUSTOMER SEGMENTATION AND TARGETED MARKETING IN MACHINE LEARNING

4.1. Machine Learning Types and Algorithms

In recent years, machine learning (ML) has transformed many industries by enabling computers to learn from data, identify patterns, and make informed decisions without human intervention. With applications ranging from image recognition to medical diagnostics and self-driving cars, machine learning represents a broad field filled with diverse methodologies. Each type of machine learning — supervised, unsupervised, semi-supervised, and reinforcement learning — has unique characteristics and applications (Figure 3).

4.1.1. Supervised Learning

Supervised learning, or learning with labeled data, is one of the most commonly used machine learning techniques. In supervised learning, a model is trained on a labeled dataset, meaning that each training example is paired with an output label. The model learns the mapping from inputs to outputs by minimizing the error between its predictions and the actual labels [4], enabling it to make accurate predictions on new, unseen data.

Classificationis a task in supervised learning where the goal is to predict a discrete class label. Classification algorithms are widely applied in fields such as targeted marketing, spam detection, sentiment analysis, and medical diagnostics. Key algorithms include:

- **Naïve Bayes:** Based on Bayes' Theorem, this simple yet effective algorithm is ideal for tasks with independent features, such as text classification. It performs well with large datasets and is highly interpretable.
- **Logistic Regression:** Despite its name, logistic regression is a classification algorithm, primarily used for binary classification problems. It predicts the probability that an instance belongs to a specific class, and its simplicity makes it a popular choice for applications like customer churn prediction.
- **K-Nearest Neighbors (KNN):** A non-parametric, instance-based learning algorithm that classifies data points based on the labels of their closest neighbors. It is particularly useful for smaller datasets and intuitive applications, such as image recognition.
- **Random Forest:** An ensemble method that combines multiple decision trees to create a more robust model, reducing the likelihood of overfitting.

Regression involves predicting a continuous output and is used in cases where the dependent variable is a real number, such as predicting house prices or stock values. Examples include:

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Random Forests:** Commonly used in many industries for both classification and regression tasks due to their accuracy and resilience to noisy data.
- **Support Vector Machines (SVM):** These powerful classifiers aim to find a hyperplane that maximizes the

margin between two classes. They perform well in high-dimensional spaces and are widely used in text and image classification.

- **Decision Trees:** Highly interpretable and effective for basic tasks, but they may overfit on complex datasets. They are also the basis for ensemble methods like random forests.
- **Simple Linear Regression:** The foundation of regression analysis, simple linear regression models the relationship between a single independent variable and the dependent variable. It is particularly useful for trend analysis.
- **Multivariate Regression:** Extends simple regression by modeling multiple features simultaneously, which is useful when multiple factors contribute to an outcome.
- **Lasso Regression:** By adding a penalty for large coefficients, Lasso Regression effectively reduces less impactful variables to zero, leading to a simpler, more interpretable model. This makes it ideal for feature selection, particularly in datasets with high dimensionality.

4.1.2. Unsupervised Learning

Unsupervised learning algorithms work with unlabeled data, meaning the model has no prior knowledge of the correct output. The goal of unsupervised learning is to uncover hidden patterns or structures within the data. Applications include customer segmentation, anomaly detection, and dimensionality reduction.

Clustering is a central task in unsupervised learning that aims to group data points into clusters such that points in the same cluster are more similar to each other than to points in other clusters:

- **K-Means Clustering:** One of the most common clustering methods, K-means works by assigning each point to one of Kclusters based on the closest mean. Although simple, K-means is highly effective for applications like customer segmentation.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A clustering algorithm well-suited for identifying clusters of arbitrary shape and handling noise. It is extensively used in spatial data analysis and image segmentation.
- **Principal Component Analysis (PCA):** A dimensionality reduction technique that transforms data into a set of linearly uncorrelated variables, or principal components. It is often used to simplify complex datasets, such as in face recognition.
- **Independent Component Analysis (ICA):** Often applied in signal processing to separate independent signals from mixed sources, such as isolating individual voices in a crowded room.

Association Rule Learning identifies relationships or patterns in large datasets. It is often applied in market basket analysis to discover products frequently bought together:

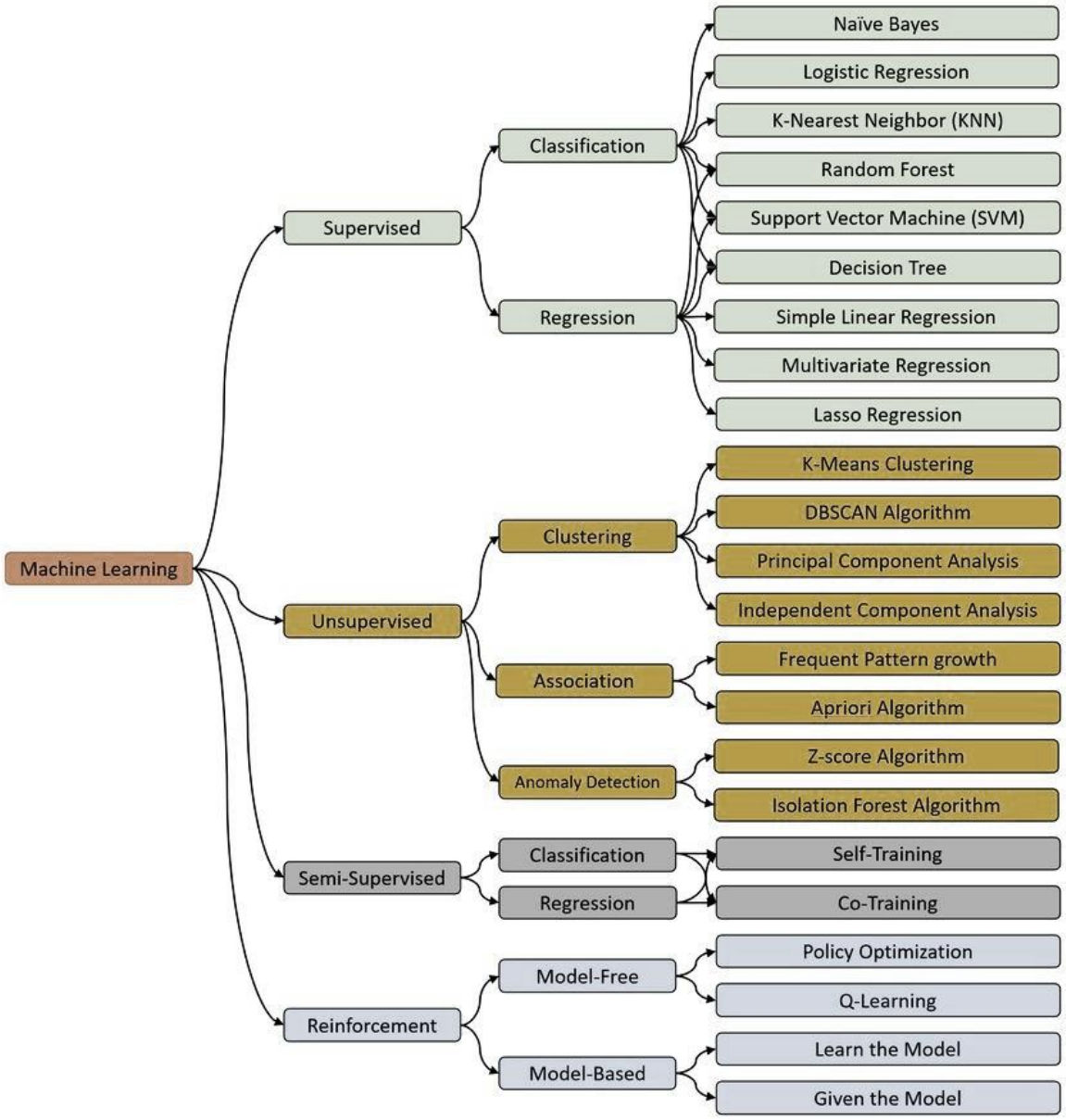
- **Frequent Pattern Growth (FP-Growth):** This algorithm improves upon traditional methods by building a compact data structure called an FP-tree, enabling efficient discovery of frequent item sets in a dataset.
- **A priori Algorithm:** A classical algorithm in association rule mining, A priori is widely used on transactional datasets. It identifies frequent item sets and generates association rules, such as discovering that customers who buy bread also tend to buy butter.

Anomaly Detection is used to identify unusual patterns that deviate from the majority of the data. It has applications in fraud detection, cybersecurity, and health monitoring:

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Z-Score Method:** Flags anomalies by measuring how far data points are from the mean, assuming a normal distribution. It is simple but effective for detecting outliers in smaller datasets.
- **Isolation Forests:** Tree-based methods specifically designed for anomaly detection. By isolating anomalies in fewer steps than regular data points, they provide a highly effective way to identify outliers in high-dimensional data.



(a)

Figure 3: Machine Learning Types and Algorithms

4.1.3. Semi-Supervised Learning

Semi-supervised learning lies between supervised and unsupervised learning, utilizing both labeled and unlabeled data. This approach is particularly advantageous when labeled data is scarce or expensive to obtain, while unlabeled data is abundant.

**Self-Training: ** In self-training, a model is first trained on a small amount of labeled data. It then predicts labels

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

for unlabeled data, and confident predictions are added to the training set. This iterative process improves the model's performance and helps it learn more complex patterns.

Co-Training: Co-training involves training two models on distinct subsets of features. The predictions from each model are then used to label the unlabeled data for the other model. This approach is useful for text categorization and other applications where data can be split into well-defined feature sets.

Semi-supervised learning has applications in fields where obtaining labeled data is challenging, such as medical image analysis, where only a few images have been annotated by specialists.

4.1.4. Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties, which it uses to improve its subsequent actions. Reinforcement learning is widely used in robotics, gaming, and autonomous systems.

**Model-Free Reinforcement Learning: ** In model-free reinforcement learning, the agent does not have a model of the environment. Instead, it learns optimal policies directly through trial and error.

- **Policy Optimization:** Policy optimization methods directly adjust the agent's policy to maximize the expected reward. These methods are often used in complex decision-making problems, such as playing games or controlling robots.
- **Q-Learning:** Q-learning is a value-based RL algorithm that seeks to learn the value of taking a specific action in a given state. This algorithm is foundational in RL and has been applied successfully in games like Atari, where agents learn to execute complex strategies by maximizing points.

**Model-Based Reinforcement Learning: ** Model-based reinforcement learning methods assume that a model of the environment exists or can be learned. This approach helps the agent to plan effective actions before making decisions.

- **Learning the Model:** The agent first learns a model of the environment, which it then uses to plan and evaluate its actions. This technique is useful in controlled environments where gathering data is costly.
- **Given the Model:** In some applications, the model of the environment is already known, such as in well-defined games. The agent can focus on policy optimization without needing to learn the model, making it efficient for such tasks.

Machine learning encompasses a diverse range of techniques and algorithms, each with its unique strengths, applications, and limitations. Understanding the distinctions between supervised, unsupervised, semi-supervised, and reinforcement learning is essential for selecting the right approach for a given problem.

4.2. Machine Learning Use Cases

MACHINE LEARNING APPLICATIONS AND USE CASES

Machine learning (ML) has become a cornerstone of modern-day technology, powering innovations that enhance decision-making, optimize processes, and enable new possibilities. From predictive analytics to real-time decision-making, ML is transforming industries with its diverse applications.

Below, we delve into some outstanding use cases categorized by learning types and their impact across various domains (Figure 4).

5.1. Supervised Learning: Teaching Machines with Labeled Data

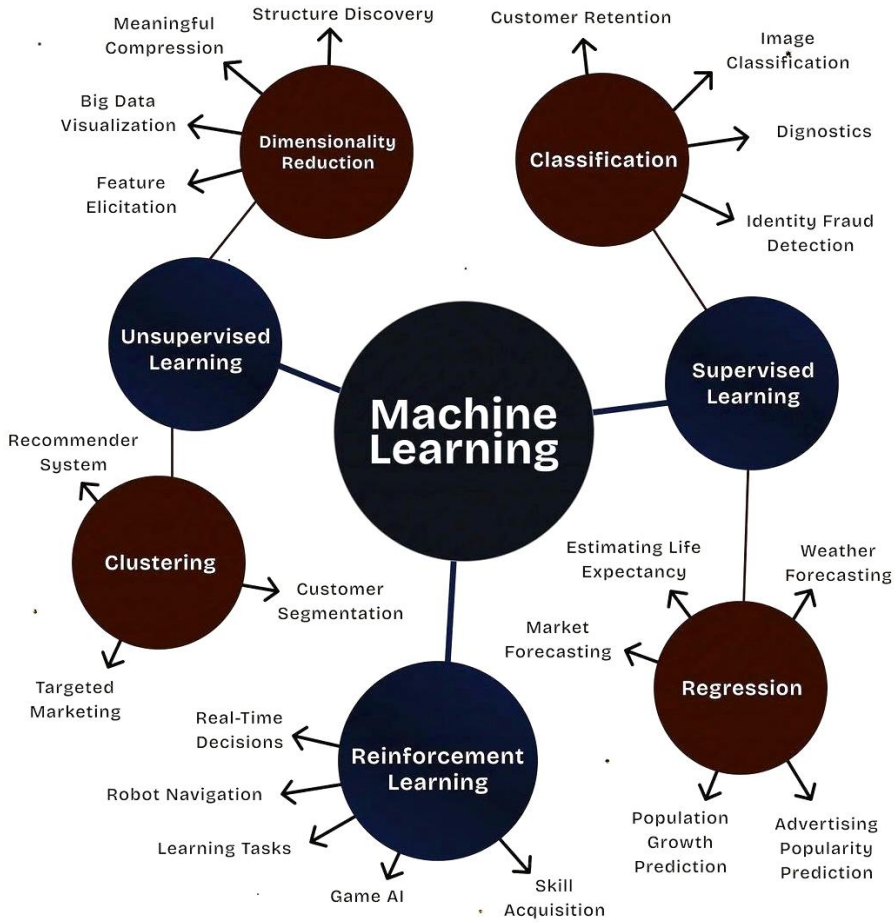
Supervised learning relies on labeled datasets to train models capable of making predictions or classifications. This approach is central to tasks requiring accuracy and reliability.

Applications of Supervised Learning

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Weather Forecasting**: ML models utilize historical data to predict emerging patterns, aiding agriculture, disaster management, and transportation.



(a)

Figure 4: Machine Learning Use Cases

- **Market Forecasting**: By identifying trends and seasonality in financial data, supervised learning enables businesses to forecast market shifts and make informed decisions.
- **Targeted Marketing**: Personalized advertisements are created by analyzing customer behavior, ensuring marketing campaigns reach the right audience.
- **Population Growth Prediction**: Governments and organizations leverage ML to understand demographic changes, assisting in urban planning and resource allocation.
- **Diagnostics and Health Monitoring**: Predictive models aid in diagnosing diseases or anomalies, providing quicker and more accurate medical insights.

5.2. Unsupervised Learning: Finding Structure in Chaos

Unsupervised learning works without labeled data, focusing on uncovering hidden patterns or relationships. This method is particularly effective in exploratory analysis.

Applications of Unsupervised Learning

- **Clustering for Customer Segmentation**: Businesses use clustering algorithms to group customers based on shared characteristics, enabling more tailored offerings.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- **Structure Discovery**: Identifying latent structures in unorganized datasets is crucial for data-driven decision- making.
- **Identity Fraud Detection**: Anomalies in transactional data are flagged using clustering methods, reducing financial fraud risks.
- **Dimensionality Reduction**: Techniques like PCA simplify complex datasets, enhancing visualization and process efficiency.

5.3. Reinforcement Learning: Learning Through Interaction

Reinforcement learning (RL) trains agents to make decisions by interacting with their environment. These agents learn optimal strategies through rewards and penalties, making RL suitable for dynamic and complex tasks.

Applications of Reinforcement Learning

- **Real-Time Decision Making**: RL powers systems that need to respond instantly, such as automated trading and emergency response solutions.
- **Robot Navigation**: Robots learn to navigate environments, avoiding obstacles and optimizing routes, which is vital for logistics and exploration.
- **Game AI and Skill Acquisition**: RL enables the development of intelligent game agents and training simulations that enhance human skills.

5.3. Cross-Domain Applications: Expanding ML Horizons

Beyond formal learning types, machine learning drives innovations in various fields:

Applications of Cross-Domain Machine Learning

- **Recommender Systems**: From streaming platforms to e-commerce, ML suggests personalized content based on user preferences.
- **Advertising Popularity Prediction**: Predictive models assess factors influencing ad performance, optimizing campaign strategies.
- **Estimating Life Expectancy**: By analyzing health, demographic, and lifestyle data, ML provides insights into longevity trends.
- **Big Data Compression and Visualization**: Advanced ML algorithms transform vast datasets into meaningful visual representations, aiding in decision-making and communication.

5.4. Key ML Learning Tasks

Underpinning these applications are foundational ML tasks, including:

- **Classification and Regression**: Core supervised learning methods used in image classification, targeted marketing, and diagnostics.
- **Feature Selection**: Selecting the most relevant data features ensures effective and accurate modeling.
- **Clustering and Dimensionality Reduction**: These unsupervised tasks streamline data analysis and visualization, forming the basis for large projects.

5.5. The Future of Machine Learning Applications

Machine learning continues to evolve, driving breakthroughs in both every day and groundbreaking applications. Whether it is guiding a robot through a manufacturing floor, recommending your next favorite movie, or predicting the weather, ML is reshaping the way we solve problems and make decisions.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

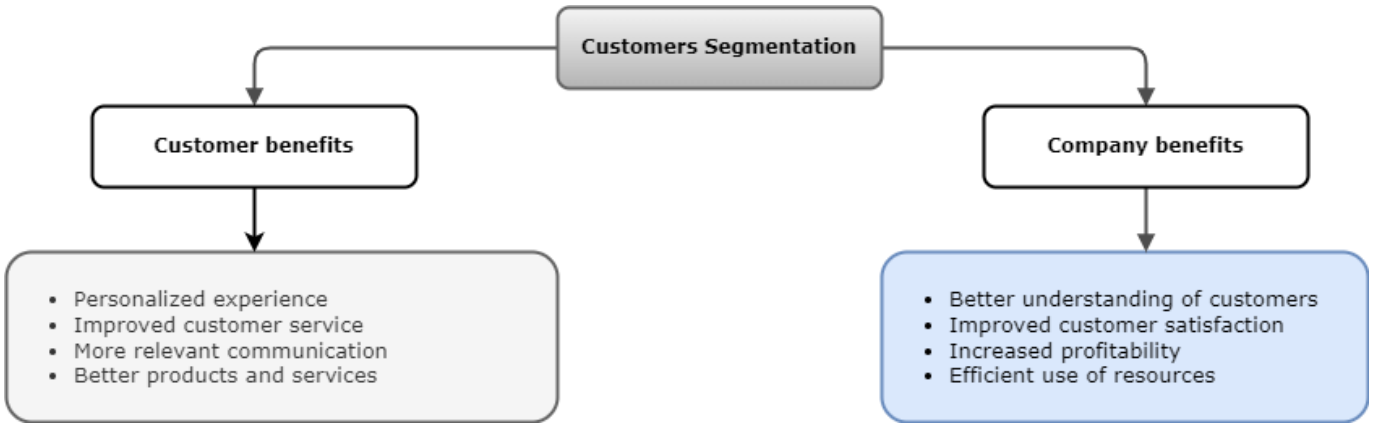
Research Article

From improving customer retention to estimating life expectancy, the possibilities are endless. As ML techniques mature and data accessibility increases, the potential for innovation grows, paving the way for smarter systems and more efficient processes.

CUSTOMER SEGMENTATION

6.1. Customer benefits

Customers can profit from customer segmentation in a number of ways. First of all, it helps businesses better understand the wants and preferences of each individual consumer, allowing them to provide a personalized experience. Targeted advertising and personalized product recommendations that are more pertinent to each individual customer can help accomplish this customization. Second, better customer service results from customer segmentation. Businesses can personalize their offerings via understanding the particular necessities of numerous client segments. For instance, businesses can deliver clients who value speed faster shipping times or supply customers who fee transparency extra thorough product statistics. Furthermore, segmentation allows organizations to interact with their customers in greater centered and applicable ways. Companies that section their consumer base can ship electronic mail campaigns and other advertising and marketing communications which can be precise to every section. This ensures that the communications resonate with clients and pique their curiosity. Finally, client segmentation delivers insights that assist agencies create higher products and services. Companies can decide the characteristics and attributes that customers fee the most by means of evaluating their preferences and behavior across many purchaser categories. This know-how permits them to enhance existing services and create new products and services that meet the specific desires of each consumer category. Customer segmentation permits corporations to offer an extra customized revel in, boom customer service, speak greater effectively, and create higher services and products. These perks help to increase patron happiness and loyalty, which drives business boom and success.



(a)

Figure 5: Taxonomy: Customer and company views.

6.2. Company benefits

Customer segmentation provides many benefits for businesses. Firstly, it allows companies to better understand their customers by grouping them into distinct categories based on demographics, behavior, or needs. This insight enables the development of tailored marketing strategies, customized product offerings, and personalized customer service that resonate with each segment. Secondly, customer segmentation leads to enhanced customer satisfaction. By tailoring products and services to meet the unique requirements of each segment, companies can improve the overall customer experience, fostering customer loyalty and positive word-of-mouth [101]. Additionally, customer segmentation contributes to increased profitability. By identifying the most lucrative customer segments, companies can focus their marketing and sales efforts more efficiently, optimizing resource allocation and generating higher returns on investment, resulting in improved revenue and profitability. Furthermore, customer segmentation enables the efficient use of resources. By directing marketing and sales efforts towards specific segments, companies can avoid unnecessary expenditures on customers unlikely to be interested, maximizing the efficiency of these activities

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

while reducing costs. In summary, customer segmentation empowers companies to better understand and serve their customers, enhancing satisfaction and loyalty, and ultimately driving business growth. Leveraging the insights gained from segmentation, companies can optimize their strategies, increase profitability, and maintain a competitive advantage.

6.3. Performance ranking of machine learning techniques

6.3.1. Data Imbalance

The effects of data imbalance in customer segmentation can go two ways. First, it can lead to skewed insights and misrepresent the behavior, needs, and preferences of your customers. Segments with more customers can dominate the analysis, making it hard to see the unique traits and actions of smaller groups. Second, when building segmentation models, data imbalance can bias the results towards larger segments. This means the model might accurately predict and describe the larger segments but struggle with smaller ones due to limited representation and fewer examples to learn from.

To tackle data imbalance in customer segmentation, you need to ensure a balanced representation of customers across all segments. Techniques like oversampling the minority segments, under sampling the majority ones, or using advanced methods like synthetic data generation can help address this issue. It's essential to maintain a representative sample from each segment to gain accurate insights and tailor marketing strategies and campaigns effectively to a diverse customer base.

6.3.2. Evaluation metrics

Evaluation metrics are indispensable in measuring the performance and effectiveness of customer segmentation models or algorithms. These metrics offer quantitative insights into how well the segments reveal underlying patterns and distinguish between different customer groups. At times, the evaluation of customer segmentation centers on its predictive performance in subsequent tasks, such as assessing its impact on customer response rates, conversion rates, sales, or other significant business outcomes.

A variety of metrics can be utilized to assess the predictive performance of customer segmentation models. These include lift, accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). Each metric serves a unique purpose in evaluating different dimensions of the segmentation's performance.

Accuracy, for instance, represents the overall correctness of the predictions and is determined by dividing the number of correctly classifiedinstances by the total number of instances. It is computed using the following equation (eq1), derived from the confusion matrix.

$$ACC = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

where TP denotes true positive, TN stands for true negative, FP means false positive, FN denotes false negative. Precision measures the proportion of true positives out of the predicted positives, providing insights into the accuracy of positive predictions It is calculated using the following equation (eq2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall, on the other hand, measures the proportion of true positives out of the actual positives, indicating how well the segmentation captures all positive instances It is calculated using the following equation (eq3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

The F1 score is the harmonic mean of precision and recall, offering a balanced assessment of both metrics It is calculated using the following equation (eq4):

$$F1score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The AUC is a metric that quantifies the performance of models in distinguishing between positive and negative classes.

It is calculated by measuring the area under the ROC curve, which represents the trade-off between true positive rate and false positive rate. A higher AUC score indicates a stronger discriminatory power of the model AUC [97] can be expressed as the following formula (eq5):

$$AUC = \frac{1+\frac{TP}{TP+FN}-\frac{FP}{FP+TN}}{2} \quad (5)$$

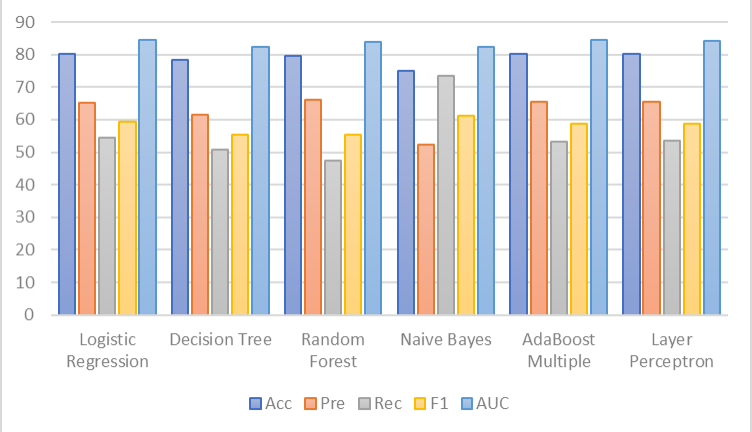
In this review, we focus on those metrics as the main metrics for evaluating the performance of customer segmentation.

6.3.3. Ranking of techniques

In this section, we analyze and compare the performances of common techniques found in related literature. We evaluate various classification algorithms applied to the Cell2Cell dataset (Table 1), provided by the Teradata Center for Customer Relationship Management at Duke University.

- Accuracy (Acc): Logistic Regression achieved the highest accuracy score of 80.19%, closely followed by Layer Perceptron at 80.12% and AdaBoost Multiple at 80.08%. Random Forest, Decision Tree, and Naive Bayes had slightly lower accuracy scores.
- Precision (Pre): Random Forest achieved the highest precision score of 66.10%, followed by Logistic Regression at 65.17% and Layer Perceptron at 65.46%. Naive Bayes had the lowest precision score among the listed algorithms.
- Recall (Rec): Naive Bayes achieved the highest recall score of 73.51%, indicating its ability to correctly identify positive instances. Logistic Regression, Layer Perceptron, and AdaBoost Multiple had relatively lower recall scores ranging from 53.24% to 54.57%.
- F1 Score (F1): Naive Bayes achieved the highest F1 score of 61.12%, effectively balancing precision and recall. Logistic Regression, Layer Perceptron, and AdaBoost Multiple had slightly lower F1 scores, ranging from 58.61% to 59.37%.
- AUC: AdaBoost Multiple achieved the highest AUC score of 84.51%, indicating its strong ability to distinguish between positive and negative instances. Logistic Regression and Layer Perceptron also performed well, with AUC scores ranging from 84.06% to 84.36%.

Based on these metrics (Figure 6), Logistic Regression, Layer Perceptron, and AdaBoost Multiple stand out as the top-performing algorithms for the Cell2Cell dataset. They exhibit high levels of accuracy, precision, recall, F1 scores, and AUC.



Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 6: The accuracy from Cell2Cell dataset

In the second part, we evaluated other algorithms applied to a bank customer dataset and examined their performance using different metrics.

Table 1. Rank from Cell2Cell dataset

Methods	Acc	Pre	Rec	F1	AUC	Rank ACC	Rank AUC
Logistic Regression	80.19	65.17	54.57	59.4	84.36	1	2
Layer Perceptron	80.12	65.46	53.4	58.8	84.06	2	3
AdaBoost Multiple	80.08	65.39	53.24	58.6	84.51	3	1
Random Forest	79.55	66.1	47.51	55.3	83.79	4	4
Decision Tree	78.33	61.5	50.72	55.3	82.37	5	5
Naive Bayes	75.14	52.32	73.51	61.1	82.3	6	6

- The Artificial Neural Network (ANN) demonstrated high precision, recall, and F-measure scores, indicating its accurate and comprehensive classification performance. With a recall score of 100%, it correctly identified all positive instances. Additionally, the AUC score of 98% indicates its strong ability to distinguish between positive and negative instances.
- The Deep Neural Network (DNN) achieved high precision, recall, and F-measure scores, indicating accurate and comprehensive classification performance. The balanced precision and recall scores suggest a good trade-off between correctly identifying positive instances and minimizing false positives. Moreover, the AUC score of 97% suggests its strong discriminatory power in distinguishing between positive and negative instances.
- The Support Vector Machine (SVM) achieved a relatively lower precision score compared to the other methods. However, it had a perfect recall, correctly identifying all positive instances. The F-measure and AUC scores suggest that it exhibited reasonably good overall performance by balancing precision and recall while effectively distinguishing between positive and negative instances.

Based on these metrics (Figure7), all three methods (ANN, SVM, and DNN) demonstrated strong performance on the bank customer dataset, exhibiting high precision, recall, F-measure, and AUC scores.

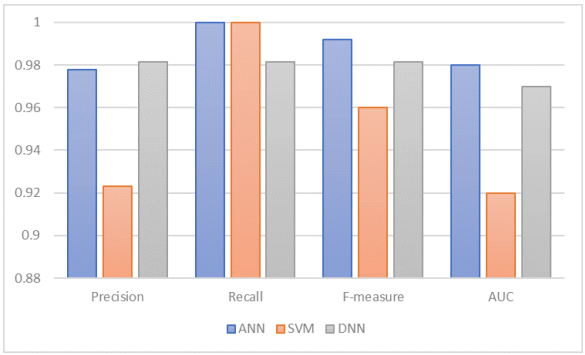


Figure 7: The accuracy from bank customer dataset

Table 2

Source	XGBOOST	Kmean	SOM	LCA	RF	FRBS	DT	RT	Cart	MOB	DBN	SVM	DBSCAN	CTREE	BMCDT	RBDT

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

[111]	+				+	+										
[66]		+					+									
[75]	+															
[16]					+											
[38]								+								
[76]			+						+	+						
[67]			+						+							
[85]		+										+				
[53]			+													
[29]			+													
[88]		+			+								+			
[11]					+						+			+		
[6]			+													
[6]			+							+						
[31]												+				
[112]	+															
[32]																
[79]					+			+								
[60]												+				
[9]												+				
[48]				+												
[8]								+							+	+
[39, 99, 96]					+											
[43, 54, 50, 46] [5, 2, 89, 83] [70, 72, 22]								+								
[91, 108, 45, 37]																
[30, 49, 14, 58]		+														
[61, 103, 71]																
[63, 15, 69, 98]																
[52, 106, 12]																

7. Results and Discussions

7.1. The Summary Tables

Our survey findings are presented in Table 7.1, Table 7.1 and Table 2. From Table 7.1, it is evident that statistical learning techniques heavily rely on LR models. On the other hand, machine learning techniques primarily employ kmean, Tree-related models, XGBOOST and SOM models, as evident from table 2. However, CART, CTREE, BMCDT and RBDT have received relatively low utilization rates. Table 7.1 illustrates that deep learning techniques, especially RNN, ANN, CNN and DNN models, are widely used. Additionally, Table 3 highlights several significant works that employ unique datasets and customized computing models to enhance the original algorithms or to evaluate their models.

Papers containing statistical models

Source	Bayesian	LR	LDA	HOSVD	PLSA
[39]		+			
[26]			+		
[77]			+	+	
[110]					+
[90]		+			
[6]				+	
[79]		+			
[32]	+				

Papers containing deep learning models

Source	RNN	ANN	DNN	CNN	LSTM
[16]	+				
[78]		+			
[92]	+				
[73]			+		
[74]	+			+	+
[102]				+	
[31]		+	+		

7.2. The datasets used in reviewed articles

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Papers under review uses a variety of datasets of customer segmentation. The type of dataset used in customer segmentation research depends on the research objectives and the available data. Figure 8 shows the distribution of customer segmentation application domain.

The main research contexts detected are as:

- **Retail**: customer segmentation is widely used in retail industry since it's a valuable tool for retailers to improve their marketing strategies [5] and offerings, increase customer loyalty [2] and satisfaction, and ultimately drive.

Table 3. A summary table with datasets

Source	Data Set	Publication	Year	Citation	Models
[2]	Real dataset for online retail store from kaggle	2022	3		Kmean
[49]	201 record from native shoping center	2021	5		kmean
[38]	Data were collected from product reviews in 3 broad categories: telephone, groceries, and fashion.	2021	4		DT, RF, GB,Super Vector Machine SVM
		2021	3		kmean
[14]	Dataset from UCI Machine Learning repository	2021	27		LDA, SOM, CART Classification and Regression Tree
[76]	Data vegetarian friendly restaurants in Bangkok				ANT DT
[54]	Publicly available customer data related to digital markets	2021	15		
[85]	Customer base from a company's direct marketing campaigns	2021	12		SVM, ensemble learning, kmean
[29]	Starting from a dataset of recipes	2021	9		self-organizing map SOM
[103]	3 different data set [88]	2021	24		Kmean clustering
334,641 transaction data					RFM model with Kmean, KMedoids, and DBSCAN algorithms
[71]	Real data set obtained from Central Insurance Company in Iran	2020	34		kmean RF,
		2020	8		LDA
[99]	Real-world dataset	2020	7		kmean
[69]	100-pattern two-factor dataset derived from the retail trade	2020	11		BMCDT, RBDT, BBRT
		2020	26		kmean
[8]	Real-time energy consumption dataset	2020	11		Kmean
[98]	Commercial International Bank of Egypt				SOM
[5]	Online retail dataset from kaggle	2020	9		SOM, HOSVD, CART
[6]	Canary Islands hotels dataset	2019	166		decision tree DT
[6]	Collection of datasets from travelers' ratings and textual reviews of spa hotels on several features in TripAdvisor	2019	222		Markov Chain Model with Decision Tree Learning DT
[50]	Two datasets of insurance and telecommunication Six	2018	72		ANN, SVN, DNN
online store datasets with annual revenues in the order of tens of millions of euros for the comparison [46]		2018	41		SVM, fuzzy match
[31]	German credit dataset to evaluate the customer retention	2018	6		Support vector clustering SVC
Data from three different Twitter business account owners [60]		2016	49		
		2015	10		
[9]	Sample from consumers in Brazil about mobile TV service				

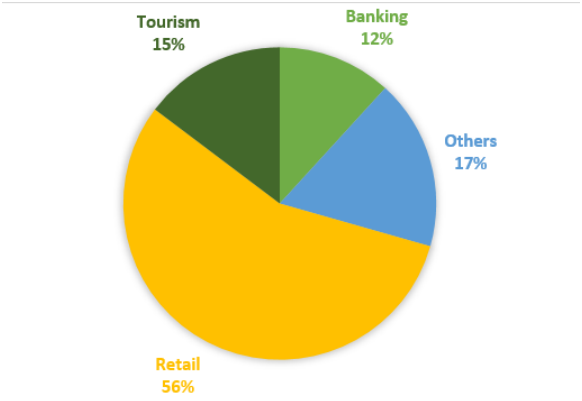


Figure 8: Distribution of dataset used in customer segmentation research

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

sales and profitability. The datasets cover different markets such as trading [95], groceries [38], fashion [38], cosmetics [86], consumer electronics [38, 86, 57], and wine [65, 74], and include data on customer behavior [64], purchase patterns [88][110], and preferences[57]. The types of data collected in these datasets vary from sales data and transaction data to product reviews [74] and direct marketing campaign data [54] [85].

- **Tourism**: Customer segmentation is important in the tourism sector to better understand the needs, preferences, and behaviors of different types of travelers. Datasets used includes different types of customer data in the tourism industry, including hotel customer behavior data, traveler reviews [6], and ratings [7]. These datasets provide a wide range of variables such as demographic, geographic, psychographic, and behavioral factors [10], that can be used for customer segmentation and analysis. The datasets also cover different regions, such as the Czech Republic [78], Malaysia [77], and the Canary Islands [7], which can provide valuable insights into customer behavior and preferences in different geographic locations.
- **Banking**: customer segmentation in the banking industry can help banks improve customer satisfaction and retention, increase cross-selling opportunities, and ultimately drive profitability. By tailoring their services and offerings to the needs and preferences of different customer segments like in studies [98], [1] and [100], banks can build stronger relationships with their customers and improve their overall financial performance.
- **Others**: the remaining reviewed papers discussed customer segmentation across various industries, including energy [8, 105], telecom [84], and insurance []. In these industries, customer segmentation is used to better understand customers and tailor marketing and service offerings to their specific needs and preferences.

The variety of used datasets provides a comprehensive view of the customer landscape in the different industries, allowing for targeted marketing campaigns, customized offerings, and improved customer satisfaction and loyalty.

7.3. Customer segmentation taxonomy in reviewed articles

Reviewed articles about customer segmentation can be categorized based on whether the segmentation approach is focused on customer profit or enterprise profit.

- **Customer-centric segmentation**: This type of segmentation focuses on understanding and categorizing customers based on their needs [89,70,49], behaviors [110,102,79,35], and preferences [26,89,1,44]. The aim is to provide a personalized and tailored experience for each customer, which in turn leads to increased customer satisfaction and loyalty. Customer-centric segmentation is often based on data analysis, such as demographic, psychographic, and behavioral data.
- **Profit-focused segmentation**: This type of segmentation focuses on categorizing customers based on their potential profitability to the company. The aim is to identify high-value customers who are likely to generate the most revenue for the company and provide them with incentives and offers to increase their loyalty and spending. Some common methods used in Profit-focused segmentation include RFM (recency, frequency, monetary value) analysis [111, 83, 40, 22, 91], customer lifetime value (CLV) modeling [16, 63, 62, 46], and profitability analysis [85, 29, 36, 68]. By combining these methods, Businesses can identify valuable client segments and build tactics to boost profitability and customer loyalty within each category.

This type of studies includes also Churn detection [72, 91, 18, 14, 32]. Customer churn can be costly for a company as it can result in lost revenue, decreased profitability, and a negative impact on the company's reputation. Therefore, companies need to focus on retaining their existing customers and preventing churn [103, 90, 31].

It's worth noting that these two categories are not mutually exclusive and can be used in conjunction with each other. For example, a company could use customer-centric segmentation to identify customer needs and preferences and then use profit-focused segmentation to provide targeted offers to high-value customers based on those needs and preferences.

7.4. Challenges

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

As with any research field, machine learning-driven customer segmentation comes with its own variety of difficulties. Some of the key challenges in this area are Data Quality and Availability, Algorithm Selection, Feature Engineering, Business Context and Ethical and Legal Considerations. Machine learning algorithms rely heavily on data, and the quality and availability of data can significantly impact the accuracy and effectiveness of customer segmentation [94]. Challenges may arise in obtaining clean, accurate, and comprehensive data that is representative of the target customer population. Data may also be fragmented or siloed, making it challenging to integrate and analyze for segmentation purposes. There are numerous machine learning algorithms available for customer segmentation, each with its own strengths and weaknesses. Selecting the appropriate algorithm for a specific business problem can be challenging [104], as it requires careful consideration of factors such as data size, data complexity, interpretability, and desired segmentation outcomes. Furthermore, some machine learning algorithms may be difficult to read, making it challenging to explain and comprehend the reasoning behind the segmentation results. Identifying the most relevant features or variables for segmentation can be challenging, as not all variables may be informative or relevant for different customer groups [93]. Feature engineering, which involves selecting, transforming, and combining features, can significantly impact the accuracy and interpretability of the segmentation results. Moreover, dealing with high-dimensional data

can be challenging, as it may result in increased computational complexity and overfitting. Overfitting is a typical problem in machine learning driven consumer segmentation, in which a model performs well on training data but fails to generalize to unseen data [25]. Overfitting can lead to inaccurate and biased segmentation results, as the model may learn patterns from noise or irrelevant information in the training data. Ensuring that the segmentation model generalizes well to unseen data and avoids over fitting is a critical challenge. Customer segmentation is often conducted to support strategic business decisions. Therefore, it is important to interpret the results in the context of the business problem and make actionable recommendations [13]. Interpreting and translating the segmentation results into meaningful insights that can guide marketing strategies, product development, and other business decisions can be challenging, as it requires a deep understanding of both the machine learning techniques and the business context. Machine learning-driven customer segmentation may raise ethical and legal concerns related to privacy, bias, fairness, and transparency [47]. Ensuring that customer segmentation is conducted in a fair and ethical manner, and complies with relevant laws and regulations, such as data protection and anti-discrimination laws, is a significant challenge that researchers need to address. Implementing machine learning-driven customer segmentation in a real-world business environment can be challenging, as it may require changes in organizational processes, systems, and culture. Managing change, addressing resistance, and ensuring smooth implementation of the segmentation results into the business operations can be a complex and ongoing challenge.

CONCLUSIONS

In conclusion, machine learning-driven customer segmentation has emerged as a pivotal strategy for organizations seeking to better understand and cater to customer needs. This study underscores the widespread adoption of machine learning algorithms, including K-means, Decision Trees, Random Forests, XGBoost, and Self-Organizing Maps, across various industries such as retail, finance, and tourism. These methods offer significant advantages in terms of precision and scalability, enabling businesses to derive actionable insights that inform targeted marketing strategies and improve customer satisfaction.

Despite its potential, the effectiveness of machine learning-based segmentation hinges on the quality and availability of data, the choice of appropriate algorithms, and the application of robust feature engineering techniques. Ensuring high-quality data and selecting meaningful features are critical for improving both the accuracy and interpretability of segmentation results. Furthermore, translating these results into actionable business insights requires an inter disciplinary understanding of machine learning techniques and the specific business context.

Future research must address challenges related to data privacy, bias, fairness, and transparency to ensure that segmentation practices align with ethical and legal standards, including data protection and anti-discrimination laws. Establishing frameworks for responsible AI and promoting transparency in segmentation processes will be key to building trust among stakeholders. By tackling these challenges, machine learning-driven customer segmentation

can continue to evolve as a powerful tool for enhancing marketing efficacy and delivering value to both businesses and consumers.

REFERENCES

[1] Abbas, Z., Merbis, R., Motruk, A., 2020. Leveraging machine learning to deepen customer insight. Applied Marketing Analytics 5.
[2] Abidar, L., Asri, I.E., Zaidouni, D., Ennouaary, A., 2022. A data mining system for enhancing profit growth based on rfm and clv. 2022 9th International Conference on Future Internet of Things and Cloud (FiCloud) , 247–253URL:https://ieeexplore.ieee.org/document/ 9910557 , doi:10.1109/FiCloud57274.2022.00041 .
[3] Abidar, L., El Asri, I., Zaidouni, D., En-Nouaary, A., 2024. Evaluating Customer Segmentation Efficiency via Sentiment Analysis: An E-Commerce Case Study. Springer Nature Switzerland, Cham. chapter pp 223–234. pp. 223–234. URL: https://doi.org/10.1007/ 978-3-031-65038-3_18 , doi:10.1007/978-3-031-65038-3_18 .
[4] ABIDAR, L., ZAIDOUNI, D., ASRI, I.E., ENNOUAARY, A., 2023. Predicting customer segment changes to enhance customer retention: A case study for online retail using machine learning. International Journal of Advanced Computer Science and Applications 14. URL: http://dx.doi.org/10.14569/IJACSA.2023.0140799 , doi:10.14569/IJACSA.2023.0140799 .
[5] Abidar, L., Zaidouni, D., Ennouaary, A., 2020. Customer segmentation with machine learning: New strategy for targeted actions, in: Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, Association for Computing Machinery, New York, NY, USA. pp. 1–6. URL: https://doi.org/10.1145/3419604.3419794 , doi:10.1145/3419604.3419794 .
[6] Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., Weaven, S., 2019a. Market segmentation and travel choice prediction in spa hotels through TripAdvisor's online reviews. International Journal of Hospitality Management 80. doi:10.1016/j.ijhm.2019.01.003 .
[7] Ahani, A., Nilashi, M., Yadegaridehkordi, E., Sanzogni, L., Tarik, A.R., Knox, K., Samad, S., Ibrahim, O., 2019b. Revealing customers' satisfaction and preferences through online review analysis: The case of canary islands hotels. Journal of Retailing and Consumer Services 51. doi:10.1016/j.jretconser.2019.06.014 .
[8] Ahmad,T.,Huanxin,C.,Zhang,D.,Zhang,H.,2020. Smartenergyforecastingstrategywithfourmachinelearningmodelsforclimate-sensitive and non-climate sensitive conditions. Energy 198. doi:10.1016/j.energy.2020.117283 .
[9] Albuquerque, P., Alfinito, S., Torres, C.V., 2015. Support vector clustering for customer segmentation on mobile TV service. Communications in Statistics: Simulation and Computation 44. doi:10.1080/03610918.2013.794289 .
[10] Antonio, N., de Almeida, A., Nunes, L., 2020. A hotel's customers personal, behavioral, demographic, and geographic dataset from lisbon, portugal (2015–2018). Data in Brief 33. doi:10.1016/j.dib.2020.106583 .
[11] Arevalillo, J.M., 2019. A machine learning approach to assess price sensitivity with application to automobile loan segmentation. Applied Soft Computing Journal 76. doi:10.1016/j.asoc.2018.12.012 .
[12] Assunção, F., Levi, M., Furtado, P., 2015. Comparing SQL and NoSQL approaches for clustering over big data. International Journal of Business Process Integration and Management 7. doi:10.1504/IJBPIM.2015.073657.
[13] Avramova, V., Smith-Miles, K., Shanahan, M., 2018. A review of business contexts for machine learning. Data Science and Engineering 3, 165–183.
[14] Bagul, N., Berad, P., Surana, P., Khachane, C., 2021. Retail customer churn analysis using RFM model and k-means clustering. International Journal of Engineering Research & Technology (Ijert) 10.
[15] Bansal, A., Shukla, A., 2020. Online insurance business analytics approach for customer segmentation. International Journal of Advance Science and Technology 29.
[16] Bauer, J., Jannach, D., 2021-06. Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. ACM Transactions on Knowledge Discovery from Data 15. doi:10.1145/3441444 . publisher: Association for Computing Machinery.
[17] Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53, 370–418.
[18] Binh, T.V., Thy, N.G., Phuong, H.T.N., 2021. Measure of CLV toward market segmentation approach in the telecommunication sector (vietnam). SAGE Open 11. doi:10.1177/21582440211021584 .
[19] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.
[20] Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classificationand regression trees. Wadsworth .
[21] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[22] Christy, A.J., Umamakeswari, A., Priyatharsini, L., Neyaa, A., 2021-12. RFM ranking – an effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences 33, 1251–1257. doi:10.1016/j.jksuci.2018.09.004 . publisher: King Saud bin Abdulaziz University.
[23] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.
[24] Cox, D.R., 1958. The use of a concomitant variable in selecting an experimental design. Biometrics 14, 375–386.
[25] Craven, J.W., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S., 1996. Overfittingin machine learning: What it is and how to prevent it, in: Proceedings of the 1996 International Conference on ArtificialIntelligence, AAAI Press. pp. 16–23.
[26] Darko, A.P., Liang, D., 2022. Modeling customer satisfaction through online reviews: A FlowSort group decision model under probabilistic linguistic settings. Expert Systems with Applications 195. doi:10.1016/j.eswa.2022.116649 .
[27] Das, S., Nayak, J., 2022. Customer segmentation via data mining techniques: state-of-the-art review. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021 , 489–507.
[28] De Lathauwer, L., De Moor, B., Vandewalle, J., 2001. Higher-order singular value decomposition. IEEE Transactions on Signal Processing 49, 2374–2384.
[29] De Marco, M., Fantozzi, P., Fornaro, C., Laura, L., Miloso, A., 2021. Cognitive analytics management of the customer lifetime value: an artificialneural network approach. Journal of Enterprise Information Management 34. doi:10.1108/JEIM-01-2020-0029 .
[30] Demir, M.Ö., Simonetti, B., Gök Demir, Z., 2021. Political segmentation based on pictorial preferences on social media. Quality & Quantity , 1–15.
[31] Dharwadkar, N.V., Patil, P.S., 2018. Customer retention and credit risk analysis using ANN, SVM and DNN. International Journal of Society Systems Science 10. doi:10.1504/ijsss.2018.095601 .
[32] Dullaghan, C., Rozaki, E., 2017. Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers. International Journal of Data Mining & Knowledge Management Process 7. doi:10.5121/ijdkp.2017.7102 .
[33] Elawady, N., Abou-Nasr, M., Fayed, H., Ghoniemy, S., 2017. Robust bayesian decision tree. I , 1–4URL: https://ieeexplore.ieee. org/abstract/document/7973241 , doi:10.1109/ECAI.2017.7973241 .
[34] Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd 96, 226–231.
[35] Fan, C., Xiao, F., Yan, C., 2015. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automation in Construction 50. doi:10.1016/j.autcon.2014.12.006 .
[36] Goswami, S., 2018. Onerous journey of a retail coffeeoutlet chain (café la coffee,india)-(CLC) imprints of an expansionist strategy. TSM Business Review 6.
[37] Guha, P., Echagarruga, C., Tian, E.Q., 2021. Optimising marketing strategies by customer segments and lifetime values, with a/b testing. Applied Marketing Analytics 7.
[38] Hadju, S.F.N., Jayadi, R., 2021. Sentiment analysis of indonesian e-commerce product reviews using support vector machine based term frequency inverse document frequency. Journal of Theoretical and Applied Information Technology 99.
[39] Hananto, V.R., Serdült, U., Kryssanov, V., 2022. A text segmentation approach for automated annotation of online customer reviews, based on topic modeling. Applied Sciences (Switzerland) 12. doi:10.3390/app12073412 .
[40] Heldt, R., Silveira, C.S., Luce, F.B., 2021-04. Predicting customer value per product: From RFM to RFM/p. Journal of Business Research 127, 444–453. doi:10.1016/j.jbusres.2019.05.001 . publisher: Elsevier Inc.
[41] Hinton, G.E., 2009. Deep belief networks. Scholarpedia 4, 5947.
[42] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.
[43] Huang, M.H., Rust, R.T., 2021. A strategic framework for artificialintelligence in marketing. Journal of the Academy of Marketing Sciencedoi:10.1007/s11747-020-00749-9 .
[44] Jagabathula,S.,Subramanian,L.,Venkataraman, A.,2017. A model-basedprojectiontechniqueforsegmentingcustomers. arXiv:Methodology .
[45] Jaiswal, A., 2021. Data mining approach for customer segmentation. International Journal for Research in Applied Science and Engineering Technology 9. doi:10.22214/ijraset.2021.35140.
[46] Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., Kobulsky, M., 2018-01. Modeling and application of customer lifetime value in online retail. Informatics 5. doi:10.3390/informatics5010002. publisher: MDPI Multidisciplinary Digital Publishing Institute.
[47] Jasmontaite, L., Nordio, M., Kaplan, A., 2020. Ethical and legal challenges of customer segmentation with machine learning. AI & SOCIETY 35, 1001–1014.

[48] Jiang, S., Cai, S., Olle, G.O., Qin, Z., 2015. Durable product review mining for customer segmentation. Kybernetes 44. doi: 10.1108/ K-06-2014-0117.
[49] Keerthi, K., Thirupathamma, G.L., Vijayalakshmi, N., Aparna, D., Vineela, U., 2021. Customer segmentation analysis and visualization. International Journal of ScientificResearch in Computer Science, Engineering and Information Technology doi:10.32628/cseit217144.
[50] Khalili-Damghani, K., Abdi, F., Abolmakarem, S., 2018-12. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. Applied Soft Computing Journal 73, 816–828. doi:10.1016/j.asoc.2018.09.001. publisher: Elsevier Ltd.
[51] Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological cybernetics 43, 59–69.
[52] Koolen, D., Sadat-Razavi, N., Ketter, W., 2017. Machine learning for identifying demand patterns of home energy management systems with dynamic electricity pricing. Applied Sciences (Switzerland) 7. doi:10.3390/app7111160.
[53] Kovács, T., Ko, A., Asemi, A., 2021. Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. Journal of Big Data 8. doi:10.1186/s40537-021-00529-4.
[54] Kozak, J., Kania, K., Juszczuk, P., Mitrega, M., 2021. Swarm intelligence goal-oriented approach to data-driven innovation in customer churn management. International Journal of Information Management 60. doi:10.1016/j.ijinfomgt.2021.102357.
[55] Lazarsfeld, P.F., Henry, N.W., 1968. Latent class analysis. American journal of sociology 73, 1–16.
[56] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324. doi:10.1109/5.726791.
[57] Lee, H., Park, Y., 2021. Data-driven approaches for discovery and prediction of user-preferred picture settings on smart tvs, in: ACM International Conference on Interactive Media Experiences, Association for Computing Machinery, New York, NY, USA. p. 134–143. URL: https://doi.org/10.1145/3452918.3458798, doi:10.1145/3452918.3458798.
[58] Lee, Z.J., Lee, C.Y., Chang, L.Y., Sano, N., 2021. Clustering and classificationbased on distributed automatic feature engineering for customer segmentation. Symmetry 13. doi:10.3390/sym13091557.
[59] Liberati, A., Altman, D.G., Tetzlaff,J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. BMJ 339, b2700. doi:10.1136/bmj.b2700.
[60] Lo, S.L., Chiong, R., Cornforth, D., 2016. Ranking of high-value social audiences on twitter. Decision Support Systems 85. doi:10.1016/j. dss.2016.02.010.
[61] Maghawry, A., Al-qassed, A., Awad, M., Kholief, M., 2021. Automated market analysis by RFMx encoding based customer segmentation using initial centroid selection optimized k-means clustering algorithm. IJCI. International Journal of Computers and Information 8. doi:10.21608/ijci.2021.207737.
[62] Marisa, F., Ahmad, S.S.S., Yusof, Z.I.M., Fachrudin, Aziz, T.M.A., 2019. Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using k-means clustering and LRFM model. International Journal of Integrated Engineering 11. doi:10.30880/ijie.2019.11.03.018.
[63] Marisa, F., Ahmad, S.S.S., Yusoh, Z.I.M., Akhriza, T.M., Maukar, A.L., Widodo, A.A., 2020. Analysis of relationship CLV with 8 core drives using clustering k-means and octalysis gamificationframework. Journal of Theoretical and Applied Information Technology 98.
[64] Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., Haltmeier, M., 2020-03. A machine learning framework for customer purchase prediction in the non-contractual setting. European Journal of Operational Research 281, 588–596. doi:10.1016/j.ejor.2018.04.034. publisher: Elsevier B.V.
[65] Mathew, R.M., Suguna, R., Shyamala Devi, M., 2019. Composite model fabrication of classificationwith transformed target regressor for customer segmentation using machine learning. International Journal of Engineering and Advanced Technology 8. doi:10.35940/ijeat. F8257.088619.
[66] Matsui, A., Moriwaki, D., 2022. Online-to-offlineadvertisements as field experiments. Japanese Economic Review 73. doi: 10.1007/ s42973-021-00101-y.
[67] Mehrbakhsh Nilashi Sarminah Samad, Behrouz Minae Bidgoli Fahad Mahmoud Ghabban, E.S., 2021. Online reviews analysis for customer segmentation through dimensionality reduction and deep learning techniques. Arabian Journal for Science and Engineering 46, 8697–8709.
[68] Miyan, M., 2017. Applications of data mining in banking sector. International Journal of Advanced Research in Computer Science 8.

[69] Monil, P., 2020. Customer segmentation using machine learnin. International Journal for Research in Applied Science and Engineering Technology 8. doi:10.22214/ijraset.2020.6344.
[70] Mosaddegh, A., Albadvi, A., Sepehri, M.M., Teimourpour, B., 2021. Dynamics of customer segments: A predictor of customer lifetime value. Expert Systems with Applications 172. doi:10.1016/j.eswa.2021.114606.
[71] Mousavi, S., Boroujeni, F.Z., Aryanmehr, S., 2020. Improving customer clustering by optimal selection of cluster centroids in k-means and k-medoids algorithms. Journal of Theoretical and Applied Information Technology 8.
[72] von Mutius, B., Huchzermeier, A., 2021-12. Customized targeting strategies for category coupons to maximize CLV and minimize cost. Journal of Retailing 97, 764–779. doi:10.1016/j.jretai.2021.01.004. publisher: Elsevier Ltd.
[73] Nguyen, S.P., 2021. Deep customer segmentation with applications to a vietnamese supermarkets' data. Soft Computing 25. doi:10.1007/ s00500-021-05796-0.
[74] Ni, P., Li, Y., Chang, V., 2020. Recommendation and sentiment analysis based on consumer review and rating. International Journal of Business Intelligence Research 11. doi:10.4018/ijbir.2020070102.
[75] Nilashi, M., Abumalloh, R.A., Minaei-Bidgoli, B., Abdu Zogaan, W., Alhargan, A., Mohd, S., Syed Azhar, S.N.F., Asadi, S., Samad, S., 2022. Revealing travellers' satisfaction during COVID-19 outbreak: Moderating role of service quality. Journal of Retailing and Consumer Services
[76] doi:10.1016/j.jretconser.2021.102783.
[77] Nilashi, M., Ahmadi, H., Arji, G., Alsalem, K.O., Samad, S., Ghabban, F., Alzahrani, A.O., Ahani, A., Alarood, A.A., 2021a. Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. Journal of Retailing and Consumer Services
[78] doi:10.1016/j.jretconser.2021.102630.
[79] Nilashi, M., Mardani, A., Liao, H., Ahmadi, H., Manaf, A.A., Almukadi, W., 2019. A hybrid method with TOPSIS and machine learning techniques for sustainable development of green hotels considering online reviews. Sustainability (Switzerland) 11. doi: 10.3390/ su11216013.
[80] Nilashi, M., Minaei-Bidgoli, B., Alrizq, M., Alghamdi, A., Alsulami, A.A., Samad, S., Mohd, S., 2021b. An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels. Expert Systems with Applications 186. doi: 10.1016/j.eswa.2021. 115722.
[81] Peker, S., Kocyigit, A., Eren, P.E., 2017. A hybrid approach for predicting customers' individual purchase behavior. Kybernetes 46. doi:10.1108/K-05-2017-0164.
[82] Qin, Y., Wang, C., Liu, G., 2019. Bmcdt: Boosting multi-class decision trees for multi-class classification. IEEE Access 7, 66659–66670. doi:10.1109/ACCESS.2019.2911227 .
[83] Quinlan, J.R., 1986a. Induction of decision trees. Machine Learning 1, 81–106. URL: https://doi.org/10.1023/A:1022643204877 , doi:10.1023/A:1022643204877.
[84] Quinlan, R., 1986b. The use of decision trees for pattern recognition in computational chemistry. Journal of Computational Chemistry 7, 616–623.
[85] Rachman, F.P., Santoso, H., Djajadi, A., 2021. Machine learning mini batch k-means and business intelligence utilization for credit card customer segmentation. International Journal of Advanced Computer Science and Applications 12. doi:10.14569/IJACSA.2021.0121024.
[86] Raja Abbas, H., Koobgrabe, C., Chutima, P., 2008. Customer satisfaction toward truemove customer service.
[87] Rogi , S., Kaš elan, L., 2021. Class balancing in customer segments classificationusing support vector machine rule extraction and ensemble learning. Computer Science and Information Systems 18. doi:10.2298/CSIS200530052R.
[88] Roychowdhury, S., Li, W., Alareqi, E., Pandita, A., Liu, A., Soderberg, J., 2020. Categorizing online shopping behavior from cosmetics to electronics: An analytical framework. arXiv preprint .
[89] Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.
[90] Sembiring Brahmana, R.W., Mohammed, F.A., Chairuang, K., 2020-04. Customer segmentation based on RFM model using k-means, k-medoids, and DBSCAN methods. Lontar Komputer : Jurnal Ilmiah Teknologi Informasi 11, 32–32. doi:10.24843/lkjiti.2020.v11. i01.p04 . publisher: Universitas Udayana.
[91] Sharaf Addin, E.H., Admodisastro, N., Mohd Ashri, S.N.S., Kamaruddin, A., Chong, Y.C., 2022. Customer mobile behavioral segmentation and analysis in telecom using machine learning. Applied ArtificialIntelligence 36. doi:10.1080/08839514.2021.2009223 .
[92] Shetty, P.P., Varsha, C.M., Vadone, V.D., Sarode, S., Pradeep Kumar, D., 2019. Customers churn prediction with rfm model and building a recommendation system using semi-supervised learning in retail sector. International Journal of Recent Technology and Engineering 8.

[93] Shirole, R., Salokhe, L., Jadhav, S., 2021. Customer segmentation using RFM model and k-means clustering. International Journal of Scientific Research in Science and Technology doi:10.32628/ijsrst2183118 .
[94] Simeone, A., Zeng, Y., Caggiano, A., 2021. Intelligent decision-making support system for manufacturing solution recommendation in a cloud framework. International Journal of Advanced Manufacturing Technology 112. doi:10.1007/s00170-020-06389-1 .
[95] Singh, S., Kumar, D., Kumar, S., 2020. A review on feature selection techniques in machine learning. Neural Computing and Applications 32, 10549–10571.
[96] Squire, M., Middleton, J., 2018. The importance of data quality for machine learning, in: Proceedings of the 2018 International Conference on Software Engineering and ArtificialIntelligence, ACM. pp. 47–53.
[97] Sroka, Ł., et al., 2021. The use of the k-prototypes method in the segmentation of customers of a company in the multi-level marketing. Wiadomo ci Statystyczne. The Polish Statistician 66, 44–56.
[98] Tan, K.S., Subramanian, P., 2019. Proposition of machine learning driven personalized marketing approach for e-commerce. Journal of Computational and Theoretical Nanoscience 16. doi:10.1166/jctn.2019.8319 .
[99] Tomczak, J.M., Zie˛ba, M., 2015. Classificationrestricted boltzmann machine for comprehensible credit scoring model. Expert Systems with Applications 42. doi:10.1016/j.eswa.2014.10.016 .
[100] Umuhoza, E., Ntirushwamaboko, D., Awuah, J., Birir, B., 2020. Using unsupervised machine learning techniques for behavioral-based credit card users segmentation in africa. SAIEE Africa Research Journal 111. doi:10.23919/saiee.2020.9142602 .
[101] Urbancokova, V., Kompan, M., Trebulova, Z., Bielikova, M., 2020. Behavior-based customer demography prediction in e-commerce. Journal of Electronic Commerce Research 21.
[102] Vijayalakshmi, M., Gupta, S.S., Gupta, A., 2020. Loan approval system through customer segmentation using big data analytics and machine learning. International Journal of Advanced Science and Technology 29.
[103] Wang, X., Zhao, P., Wang, G., Liu, J., 2007. Market segmentation based on customer satisfaction-loyalty links. Frontiers of Business Research in China 1, 211–221. URL: https://doi.org/10.1007/s11782-007-0013-0, doi:10.1007/s11782-007-0013-0 .
[104] Westland, J.C., Mou, J., Yin, D., 2019. Demand cycles and market segmentation in bicycle sharing. Information Processing and Management 56. doi:10.1016/j.ipm.2018.09.006 .
[105] Wu, S., Yau, W.C., Ong, T.S., Chong, S.C., 2021. Integrated churn prediction and customer segmentation framework for telco business. IEEE Access 9. doi:10.1109/ACCESS.2021.3073776 .
[106] Yang, C., Wang, F., Zhou, Z.H., 2017. Algorithm selection for machine learning: A survey. Data Science and Engineering 2, 269–283. [105] Yuan, Y., Dehghanpour, K., Bu, F., Wang, Z., 2020. A data-driven customer segmentation strategy based on contribution to system peak demand. IEEE Transactions on Power Systems
[107] Yuliari, N.P.P., Putra, I.K.G.D., Rusjayanti, N.K.D., 2015. Customer segmentation through fuzzy c-means and fuzzy RFM method. Journal of Theoretical and Applied Information Technology 78.
[108] Zadeh, L.A., 1993. Fuzzy modeling and control. IEEE Transactions on Systems, Man, and Cybernetics 23, 262–269.
[109] Zavali, M., Lacka, E., de Smedt, J., 2021. Shopping hard or hardly shopping: Revealing consumer segments using clickstream data. IEEE Transactions on Engineering Management doi:10.1109/TEM.2021.3070069 .
[110] Zeileis, A., Hornik, K., 2008. Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17, 492–514.
[111] Zhang, Q., Yamashita, H., Mikawa, K., Goto, M., 2020. Analysis of purchase history data based on a new latent class model for RFM analysis.
[112] Industrial Engineering and Management Systems 19. doi:10.7232/IEMS.2020.19.2.476 .
[113] Zhao, X., Keikhosrokiani, P., 2022. Sales prediction and product recommendation model through user behavior analytics. Computers, Materials and Continua 70. doi:10.32604/cmc.2022.019750 .
[114] Zhuang, Y., 2018. Research on e-commerce customer churn prediction based on improved value model and XG-boost algorithm. Management Science and Engineering 12.