Detecting Customer Spending Patterns and Preferences Using AI

Balaji Soundararajan

(Independent Researcher)
esribalaji@gmail.com

Abstract

In the era of AI-driven consumer insights, businesses leverage advanced machine learning (ML) and data mining techniques to transform raw customer data into actionable intelligence. We will explore the integration of AI in customer analytics, emphasizing data collection from diverse sources (transactional, clickstream, social media, and open data), preprocessing strategies, and the application of ML models such as clustering (k-means, hierarchical, DBSCAN) and classification (decision trees, SVM, logistic regression). Case studies across industries highlight the efficacy of recommendation systems, personalized marketing campaigns, and ethical considerations in data usage. By synthesizing real-time data analysis with predictive modeling, organizations can enhance customer engagement, optimize revenue, and deliver hyper-personalized experiences while addressing challenges related to algorithmic fairness and privacy.

Keywords: Customer Analytics, Machine Learning, Data Mining, Clustering Algorithms, Classification Algorithms, Personalized Marketing, Ethical AI, Recommendation Systems, Data Preprocessing

Introduction

In this age driven by consumer choice, knowing the customer inside out has always been important. But back when researchers used to deploy long surveys, or even further back to the days of pen-and-paper data collection, tracking and analyzing consumer habits and patterns in real time was not within the scope of possibility for most organizations. This has all changed with recent developments in artificial intelligence, though. Thanks to AI, it's not enough in and of itself to merely offer customer services. Companies need to identify what products and services customers will likely want, if not before they want them, at least soon after. This is where the concepts of customer cornucopia and customer value profiler come into play.

Customer cornucopia allows organizations to go from products to people. Utilizing machine learning, companies can follow a customer's digital breadcrumbs through this mountain of available data as they make purchases, leave product reviews, and generally browse the internet in a way that enhances and leads to an increase in dynamic profiling. Data mining can detect user behavior patterns by mining and analyzing a significant amount of data. Then, once an entity knows what one likes, a complex machine learning-based system can predict whether a person will enjoy another given product, movie, song, etc., based on what else they like. More traditional applications suggest potential new menu items once a new product habit is established, and all of this can be harnessed using consumer activity and browsing history, from monthly sales figures to overall ad revenues. All of this data is used in tandem to notably increase the level of customer engagement by appealing to a more dynamic array of preferences to define value through more informed decision-making.

Data Collection and Preprocessing

Data collection and preprocessing are vital stages for customer analytics. High-quality data collection is required to increase confidence in the results of the analysis. The pervasiveness of internet-based communication systems, the rapid development of internet technology, and the diverse needs of customers have led to the introduction of various types of online databases that can be accessed by customers immediately, such as transactional databases, information databases, social media, telecommunication companies' databases, and online customer feedback databases. Through these communication channels and databases, companies can access a variety of information about customers. In addition, companies invest considerable resources in various market research campaigns that produce various forms of survey data. All this information can help companies create a more profound view of their customer base. During the preprocessing phase, one of the fundamental steps is the data transformation step, which is aimed at converting the data from a unified data matrix into the format necessary for the application of data mining algorithms. After the transformation of the primary data, it is possible to use the data mining software in order to predict customer behavior. The aim of data preprocessing is to transform the raw dataset into a smaller, clean, and relevant dataset. The examples of preprocessing steps are: - Sourcing relevant customer data from a variety of data sources, such as databases, online customer feedback systems, and social media channels - Filtering and cleaning large volumes of customer data to ensure that missing or noisy data does not compromise the results of the analysis - Data transformation: it is required to convert the raw data matrix into a compact matrix. This compact matrix can be further used for analysis using DB or visualization of the customer's pattern. The row index and full matrix represent customer transactions to create the FP tree. The non-empty column represents the FP tree with a null leaf at the root level. The non-empty column represents the result of creating a branch in the FP tree.

Types of Data Sources

Digital ecosystems have facilitated the development of diverse data sources that can be used to understand customer behavior and customer engagement with brands. Typically, four types of data sources are used in customer analytics: transaction data, clickstream data, social media data, and open data. Structurally, data can be partitioned into three types: structured, semi-structured, and unstructured. At one extreme, structured data are easy to analyze but, on the other hand, these types of data don’t offer much privacy about the customer. At the other extreme, unstructured data are not easy to manage but are more private than structured data. Therefore, these types of data provide different aspects of analysis, and therefore it's best to, where possible, use multiple sources. A detailed description of each data source will be described next.

1) **Transaction data:** Data in system logs, i.e., customer account information, purpose of buying, cost per item, quantity bought, and transaction time. This type of data can be obtained under voluntary participation, i.e., the customer is free to provide or not to provide. One or more transactions may yield a purchase to profit instances.

2) **Clickstream data:** Record of activity created by visitors/products on a website that contains IP address, referral page or link, time of last visit, time started in the session, event, interest code.

3) **Social media data:** Data about people, places, organizations, and interests listed on social media that can be obtained under non-voluntary participation, that is, one cannot force a person into having an interest in something.

4) **Open data:** Data from online data marketplaces and available in web classifieds that contain country, state, city, category, title, price, season IDs, GPS latitude, GPS longitude, species identifiers, animal IDs, date found or lost, image, and sex.

Data aggregation from both voluntary and involuntary participation can be done via relevant application programming interfaces and common web scraping software that can download and extract data from almost any web page, including those with restricted access. Each of these data sources has both merits and demerits. Because each source acts as a filter, it is necessary to aggregate data from multiple sources to offer a more complete picture of the behavior and preferences of the customer. We can also download some transactional data in real-time, most of the time, social data use may offer an opportunity to respond to a customer who has taken issues about a product immediately for analysis.

Data Cleaning and Transformation

There are several key preprocessing steps that are typically performed to get data ready for analysis. Technically, the very first step is undertaking a data auditing and understanding exercise to determine what data is available. Initial output from this process should be a data profiling report which contains various descriptive statistics and graphical representations. One of the most common issues to arise is missing data, one of the simplest fixes for which is deletion. Duplicates, while not a standard method of dealing with, can also be easily addressed through removal or further examination if necessary. We can then address outliers which can be more complex to deal with. An example of a simple approach to dealing with this in the e-commerce setting is to delete very large transactions which often have nothing to do with typical customers. In small datasets, however, this is not really a common way to deal with outliers. There are many more methods of imputation which are appropriate for use across datasets of all sizes. In most analyses, we also transform the columns of the dataset into a form appropriate for machine learning purposes. There are different methods for handling categorical columns, the most common of which is to use one-hot encoding.

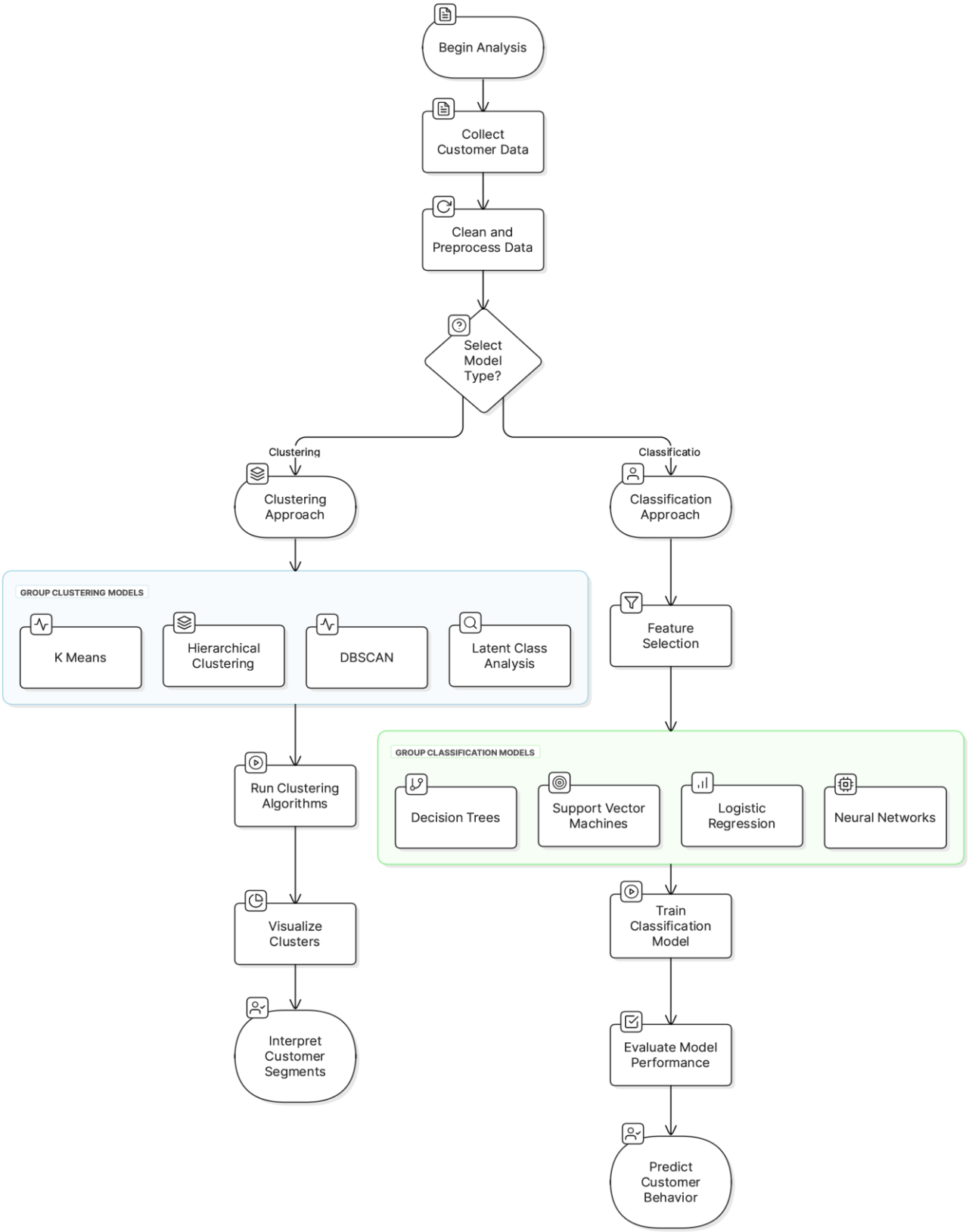
The real value of any analysis lies in the quality of the data input. Data cleaning and transformation are important steps to gain meaningful insights and predictions from the customer data. For example, if we ever wanted to build a classification model, we need to transform the data so that categorical attributes are 'dummified' or 'one-hot encoded.' This process is critical as algorithms such as random forest or k-nearest neighbors require a distance measure to determine similarity, and categorical measures can favor certain attributes based on the number of levels present. Since the inputs specify the state of the model, the quality of the data is critical to the performance of the model. [1]

Machine Learning Models for Customer Analysis

Predictive analytics allows for discovering customer preferences and spending patterns and forecasting future customer behavior. In order to do so effectively, it is necessary to develop a machine learning model that can use behavioral attributes to find relationships between transactions, wallet addresses, or features such as transaction amount, time of transaction, etc. Different types of machine learning models program and apply four main learning models to the selected problem. Given an input data point, classification is the task of assigning it to one of a predefined set of classes. Clustering groups multiple data points based on the features that the data points share with one another, with the intuition that the data points within the same cluster are more "similar" to one another than to those in other clusters. Regression, similar to classification, requires the output to be drawn from a limited number of classes, leading to a task where a continuous or discrete output is predicted based on input attributes. [2]

Clustering customer analysis techniques find customer segments with similar characteristics and shopping spending patterns based on customer behavior. The purpose of classification customer analysis techniques is to predict customer response to different marketing strategies. There are other machine-based algorithms and models like regression and association rules-based methods, data mining, statistical models, etc. that can be used. Clustering models suit retail data analytics, fraud analysis on transactional data, and targeted

advertising. Metadata-based recommendations can be found based on these models that suggest related products with history-based purchases. Classification models predict customer replies based on past responses to improve response accuracy. These models can be used to create a customer classification result and distinguish other variables and behaviors from the response. This helps in achieving a good response category. Further comparisons are done with models in open research.



Clustering Algorithms

Clustering is another unsupervised learning algorithm that is employed in the domain of customer data analysis. The idea behind clustering is to group customer data points based on similarity between the data points. This grouping would lead to identifying hidden patterns, insights, and consumer behavior. The various techniques for clustering are as follows: K-means: This is a hard clustering algorithm, where it needs the number of clusters as an input to segment the data into that number of clusters. Hierarchical clustering: There are three types of hierarchical clustering — agglomerative, divisive, and BIRCH. This clustering algorithm is based on splitting or merging the data points, which starts with each data point as its own cluster, and then at each step uses a similarity metric to merge the closest two clusters into one. DBSCAN: This is a density-based clustering method, where it can identify clusters of arbitrary shapes and sizes. Latent class analysis: This is a statistical latent variable model, which is used to identify unobserved or latent subgroups.

Clustering helps in overcoming the weaknesses of demographics and customer lifetime value techniques. It is a strategic way of understanding your target audience. Usually, after the clusters are formed, companies apply some supervised learning models to identify the features that distinguish a particular segment or group. Some of the benefits or advantages of employing clustering are personalization — segmentation of customer data can be seen as the first level of customization used in targeted marketing strategies, where the customers are targeted with offers or services that are of interest to them. The number of clusters to be used is still an open question and mainly depends on the requirement. There are many statistical methods available, like interpreting the dendrogram, to get the best number of clusters for that data set. The majority of segmentation or clustering problems primarily use the elbow method to get the optimal number of clusters. [3]

To perform customer analytics, with the help of clustering, the various steps required are as follows: The first step would be to collect all the relevant customer data. Due to the raw nature of data, one needs to clean the data by removing noise and irrelevant data. The various clustering algorithms are run one after another to identify the clusters from the data points. The results are visually displayed using graphs, like scatter, density, or a 3D plot to interpret the results.

Classification Algorithms

Classification algorithms are used to predict customer behaviors based on historical data. Modelers can build classification models and try to predict outcomes, such as who is going to churn, respond to a cross-sell, or an up-sell marketing campaign. Examples of classification algorithms include decision trees, support vector machines, neural networks, and logistic regression. Each classification technique comes with its strengths and drawbacks and might be better suited for a particular scenario. Decision trees are algorithms used for determining different outcomes based on a different set of decision rules, on the principle of making a series of choices to determine the final outcome.

Support vector machines are unique classification algorithms that find the best possible line that separates the two classes. It is the objective of the support vector machine to find lines that maximize the margin between the two separate data classes. Logistic regression is a statistical method that models the probability that a given outcome is a member of a particular category. In customer analytics, we are trying to predict an event outcome, such as whether a customer will purchase, respond to a promotion, or churn. Feature selection is a key step in building a classification model. That is because a carefully selected set of variables is going to improve the success of a classification model. Including too many features in a model that are not relevant is going to detract from the power of the model and may make it more complicated to interpret.

The quality and relevance of the features used to construct a classification model are going to have significant consequences on the predictive importance of the model. Important features used in the classification are going to provide better accuracy, while irrelevant features may lead to more, rather than fewer, errors. Variables used to identify classes and build the model explain how and why certain values vary by class. For instance, it is important in a customer profile to know how income, age, spending habits, and location differ between different groups of classes. Some metrics are used to rate the efficiency of classifiers, such as accuracy, sensitivity, specificity, precision, and F-measure. The classifier's quality can be checked by using accuracy, area under the ROC curve, etc. [1]

We can use classification algorithms in customer analytics for investigating customer preferences by analyzing the products that they are purchasing, by bio-psychological and physiological characteristics of the potential customers, which products or services they would like, or by purchasing online. With a credit card, the information on people's telecommunication and personal lives is easier to analyze. Additionally, we can use classification algorithms to investigate customer churn or why an online transaction is more likely to be fraudulent, whether you are likely to respond to a credit card offer, or whether someone is likely to buy at your store or service or through your app. Another scenario could be a combination of probability and rules, in which probability is used initially to build an initial risk profile as the basis for creating rules to manage customer risk. However, classification should not be done on a certain set of the population for which there is evidence that it would not be applicable. For example, a bad customer characteristic of known respondents may differ from the population and be sensitive to certain characteristics. Then classification would occur on the selected independent sample. There are some difficulties, though, in classification due to model overfitting and class imbalance.

Case Studies and Applications

For this section, we will provide several case studies to illustrate the different applications of AI in customer analytics across various industries. In each case, the company leverages machine learning models to better understand and predict different aspects of customers, such as spending patterns and preferences. These instances are proof that advanced customer analytics offer a set of business benefits, including an uplift in sales revenue, customer retention, and an increase in customer satisfaction through more tailored services and targeting. It is also evident that prior to implementing an AI model, the company should first define the problem they want the model to solve. Subsequently, the model's performance and business impact should be closely monitored.

The case studies include, for example:

* A company tracking the dynamics of customer opinions on its products and product assortment to create insightful and actionable reports about popular items and customer preferences.

* A business using AI solutions to better understand and predict its customer spending patterns, then offer the targeted financial product that is most likely to become profitable.

* A telecom company employing chatbots on a large scale for customer care to improve customer satisfaction and retention.

* A retail company offering an AI-based suggestion system that helps customers find items that perfectly match their preferences and get inspired when looking for fashion items.

* A bank using an AI-based product portfolio optimization model as a sales support tool, helping sellers offer customers a preselected portfolio of banking products as a cross-sell offer tailored to suit a particular customer's preferences.

E-commerce Recommendation Systems

Recommendation systems are a key functionality in e-commerce applications. They help users find products consistent with the choices of others and/or the profiles of the users. Intuitively, recommendation systems suggest that if a certain customer prefers a certain item or a set of items, the customer is likely to be interested in a similar item or in items similar to the set of items already acted upon by the customer. A well-designed and implemented recommendation system can boost sales, enrich customer visits, and provide users with extra value from their real-time data processing.

A primary strategy for e-commerce is to use three types of algorithms to send customers the best items. In addition to pushing popular items that are liked by many customers across the platform, top-k offline recommendation is recommended for accessing the high-accuracy dimensions matrix factorization and the extrapolation algorithm of customers and items to select high-quality items, as well as result merging, reranking, rescoring, and model post-processing. These activities can improve the user experience and promote strategic development. Case studies show that various platforms will reach spectacular results in a varied range of real applications, such as e-commerce, e-news, e-gaming, IoT, e-reading, e-tourism, and e-transport. However, industrial-sized recommendation systems with large-scale data and online serving require careful handling of these challenges and increasing opportunities. First, recommendation systems require secure processing of users' private data and managing recommendation algorithms that unfold several ethical and privacy-related questions. Second, recommendation systems have demonstrated how they affect user behavior and platform development in the context of increased use, drawing attention from machine learning, data science, marketing, and behavioral trust.

Personalized Marketing Campaigns

With personalized marketing, companies can adapt messages to reach consumers who spend differently. Typically, businesses use data from in-store or online spending to segment customer groups. For example, high-value customers receive marketing with discount coupons, while others may receive marketing with general information. Data analytics capabilities and advances in AI now enable companies to individualize each customer's treatment and to increase the engagement rate with a marketing message on social media. Imagine your brand creating a rapport with each individual and inviting that person to browse a unique assortment. Personalized campaigns drive a 30% conversion rate, solidify relationships with high-spending one-offs, and 24% of one-offs convert to be repeat customers.

Intelligent companies are testing and learning from campaigns by using feedback loops, A/B testing, and prioritizing the most impactful personalization actions. Consideration must be given to serve personalized campaigns responsibly and ethically. This requires a commitment to advocacy by practicing meaningful principles such as transparency and privacy with individuals to guide 'earned trust,' but also setting guardrails and embedding privacy by design. This means embedding a policy of personalization that provides the ability to adjust the technology to better deliver a foundation of privacy and security. Efficient operations depend on automating manual processes. The function of this operating process is to automate campaign production and creative personalization at scale in real time using a context-driven AI system to drive peak sales performance and optimize results.

Ethical Considerations in Customer Data Analysis

The accuracy of algorithms is heavily advertised by companies developing them. The data and the results that machine learning provides are extremely attractive. Companies have integrated machine learning into their recommendation systems, and revenue has significantly increased as a result of this effort. Many types of machine learning have been directly applied to customer data, all being used in hopes of increasing revenue. But of course, our main concern when we do data analysis is to properly use the data that we have in a correct manner. After taking these considerations into account, we can ensure that no one participating in the project can be harmed in any way. Wrong conclusions that result from our data analysis or data falsity can lead to charges against individuals and also against companies. In terms of fair decision-making, unethical data analysis concepts like accuracy and fairness of algorithms have been thoroughly researched, and we know that we have some new problems related to fairness when dealing with customer data. These studies develop methods for designing prediction algorithms with fairness. [4]

The idea of fairness in machine learning has been extensively studied. However, all methods that were developed to ensure that all individuals are treated fairly assume that comprehensive knowledge of the subgroups exists and must be determined beforehand. Many negative outcomes can occur if a person is treated more favorably than others solely based on their sensitivity and the subgroup to which they belong. Profit that can occur can be categorized according to the four different quadrants, which is captured in the two-by-two matrix. Individuals who belong to quadrant A possess a low score, meaning that they can bring profit to the company if they are approached. On the other hand, the people who belong to quadrant D possess a high score and should be approached. Then we have individuals who belong to quadrant B, who can bring the desired profit and can also be approached. Individuals in quadrant C should not be approached since they do not belong to any group of interest, so no profit can be achieved from them. Individual fairness is not being achieved by this form of treatment, and people do not form into subgroups beyond those mentioned that are not always meaningful. It is under this light that the scientific community has neglected the weak level of fairness because no one has truly studied methods to ensure a minimal level of fairness by using more conservative methods.

Conclusion:

The integration of AI in customer analytics has revolutionized how businesses understand and predict consumer behavior. By harnessing structured and unstructured data from transactional logs, social media, and clickstreams, organizations can build dynamic customer profiles through robust preprocessing and ML techniques. Clustering algorithms enable segmentation of customers into behaviorally similar groups, while classification models predict churn, preferences, and purchasing patterns. Applications such as e-commerce recommendation systems and personalized marketing campaigns demonstrate tangible benefits, including increased conversion rates and customer retention. However, ethical challengessuch as algorithmic bias and data privacyrequire proactive mitigation through transparency and fairness-aware design. Future advancements in AI must prioritize ethical frameworks, real-time adaptability, and integration with emerging technologies like generative AI to further refine customer-centric strategies.

References:

[1] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[3] F. Yoseph, N. H. Ahamed Hassain Malim, "The impact of big data market segmentation using data mining and clustering techniques," Journal of Intelligent, 2020. academia.edu

[4] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electronic Markets, 2021. springer.com