Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Optimizing Online Sales Strategies in the USA Using Machine Learning: Insights from Consumer Behavior

Rabeya Akter[1], Md Nasiruddin[2], Farhana Rahman Anonna[3], MD Rashed Mohaimin[4], Md Boktiar Nayeem[5], Adib Ahmed[6], Afrin hoque jui[7], and Shah Alam[8]

[13]Master of science in information technology. Washington University of Science and Technology, USA

[267]Department of Management Science and Quantitative Methods, Gannon University, Erie, PA, USA

[4]MBA in Business Analytics, Gannon University, Erie, PA, USA

[5]Master of Science in Business Analytics, Trine University

[8]Master of Science in Information Technology, Washington University of Science and Technology, Alexandria, VA, USA.

Corresponding Author: Rabeya Akter, **E-mail**: rakter.student@wust.com

| ABSTRACT

The exponential expansion of e-commerce in America has redefined the retail landscape, presenting opportunities and challenges for online retailers. This research aims to apply machine learning techniques to develop a strategic online sales strategy through deep consumer behavior analysis. This research paper focuses on a consumer behavior analysis based on U.S.-based datasets underlining American consumers' unique characteristics and preferences. The consumer behavior dataset contained complete data on various aspects of the customer's behavior in online retail. The dataset consisted of transaction records for customer purchase history, items purchased, frequency of purchases, and values of transactions. It also contained browsing history data that would point out user interaction patterns, such as visited pages, time spent on each page, and views of different products to draw fine-grained inferences on consumer interest and preference. The analyst implemented accredited and credible models, such as Random Forests, Logistic Regression, and Gradient Boosting Classifiers, that are useful in various dataset analyses related to customer behavior. Random Forest turned in a strong performance, having relatively high accuracy, reflecting that it is efficient in picking up complex patterns in the data. Machine learning can revolutionize the way online retailing is approached in the U.S., as it has the potential to make full use of consumer data on a large scale for more nuanced decision-making. While integrating machine learning algorithms, retailers can develop highly personalized shopping experiences that best meet the preferences and behaviors of individual customers.

| KEYWORDS

E-commerce, Consumer Behavior, Machine Learning, Personalized Marketing, U.S. Market, Online Sales Optimization, Dynamic Pricing

| ARTICLE INFORMATION

RECEIVED: 01 July 2023 **PUBLISHED:** 21 July 2023 **DOI:** 10.32996/jbms.2023.5.4.17

I. Introduction

Importance of e-commerce in the USA

Ballestar et al. (2019), reported that E-commerce has grown to be an integral part of the American economy, accounting for a huge proportion of retail sales in the nation and thereby contributing majorly toward job creation and overall economic growth. At recent estimates, e-commerce sales in the country have now crossed the trillion-dollar mark annually, with this factor besides

underlining its importance having transformed traditional retail models. Thus, it has become very significant in the American economy. According to Chaudhuri et al. (2021), The COVID-19 pandemic accelerated it even more, and consumers were bound to shift their shopping habits toward online platforms, hence a change in the landscape of consumer behavior. It is a shift that has forced online retailers to rethink their sales strategies, invest in technology, and tap into consumer data to be competitive in an ever-growing marketplace.

As per Choi & Lim (2020), the role of e-commerce in driving economic activity cannot be overstated. It helps in consumer convenience, extends business reach in the market, and promotes innovation due to technological advancement. Geographical boundaries no longer bind retailers; they can now reach customers all over the country and even the world through digital channels. This has made the competition among online retailers much fiercer, forcing them to employ advanced tools and strategies to attract and retain customers. It hence comes to a juncture whereby the understanding of consumer behavior becomes one of the most vital keys to succeeding in such an aggressive environment. The use of big data analytics and machine learning toward the simplification of such complex consumer preferences has therefore become important for every retailer (Fieldman et al., 2022).

Issues in Optimizing Online Sales

Boone et al. (2019), asserted that while e-commerce has opened doors of opportunities, the challenge to optimize the e-sales strategies of retailers remains great. One imminent challenge in this respect pertains to translating the enormous data amounts from consumer behavior, generated through digital interactions, into actionable insight. Bharadiya (2023), posited that traditional analytics tools often prove inelastic in their dissection capabilities at the complicated buying patterns that are influenced by a host of factors- Seasonality, economic conditions to individual preference. As a result, retailers may struggle to identify the most effective strategies for engaging customers and driving sales.

Furthermore, the domain of traditional analytics has a limitation when volume and speed become difficult to manage in the digital arena. While basic analytics could reveal an instant snapshot of consumer behavior, most of the time they are inadequate to uncover the subtleties that predispose people to make purchase decisions (Choi & Lim, 2020). For example, the reason for which a consumer abandons a shopping cart or what influences them to make a purchase requires deeper analysis, which may not be fully captured by traditional methods (Boppiniti, 2022).

According to Gupta et al. (2020), such gaps in understanding can lead to missed optimization opportunities and ultimately affect the bottom line of a retailer. There is also a very rapid speed of technological changes emergent tools and platforms force retailers to make continuous adjustments in their strategies. All these are in constant evolution, which requests a very proactive approach toward data analytics, which may anticipate, rather than simply react to, consumer behavior. In this respect, the application of advanced methodologies such as machine learning becomes very important.

The objective of the Research

By analyzing large, complex datasets that represent the complexity of consumer interaction with online retail platforms, actionable insights that can drive marketing strategies, product placements, and customer engagement will be uncovered. This information, obtained from the results of machine learning analyses, helps improve customer satisfaction by personalizing shopping and is likely to bring in bigger sales and higher revenue. Furthermore, this research intends to give a layout for data-driven strategies that will be adopted by retailers in coping with the continuous changes in consumer behavior. Given the focus on machine learning predictive capabilities, the present study attempts to arm the retailer with an enabling capability of being able to estimate consumer needs to customize marketing and inventory management. The use of machine learning for enabling sales comes not only with superior operational efficiency but also holds the potential for competitive advantage in a highly congested marketplace.

Scope and Relevance

This research paper focuses on a consumer behavior analysis based on U.S.-based datasets underlining American consumers' unique characteristics and preferences. The study shall, thus, seek to identify major sales drivers that may exist in the data collected from different online retail platforms. This study is of theoretical relevance but also of practical relevance for each retailer who wants a more competitive online sales strategy. These machine learning techniques in this context offer insight into consumer behavior that goes far beyond surface-level metrics to delve into the underlying motives and preferences driving consumers to make purchases. As retailers increasingly hang their strategy on the hook of data-informed insights, the insights afforded through this research can be a real ally in helping to develop targeted marketing campaigns, optimize product offerings, and generally enhance customer engagement.

II. Literature Review

E-commerce Consumer Behavior

Sharma et al. (2020), argued that the process of understanding consumer behavior in e-commerce is very complex, having dismembered various variables regarding customers' purchase decisions in the United States. It's been made very easy to purchase from wherever anytime. Research has depicted time-saving and access-easy as essential characteristics in driving online purchasing. Moreover, the sensitivity of price remains very crucial: people make price comparisons at different places for better offers. The existence of discount codes, flash sales, and seasonal promotions only heightens this trend of being price-centric.

Another important factor of online shopping, which cannot be overemphasized, is the role of personalization. Personalization not only caters to the consumer's individual preferences but also enhances their overall shopping experience. E-commerce sites using data analytics to personalize product recommendations based on purchase history, browsing, and demographics have a significant improvement in conversion rates (Khrais, 2020). People tend to make more purchases when products match their tastes. Furthermore, digital engagement for the capture and retention of customers has become so important with the usage of interactive content, personalized email marketing, and target advertisements. In an over-choice marketplace, engagement may make the difference for a brand between being just another choice and being chosen (Khrais, 2020).

Besides personalization and interactive elements, social proof has come to be a particularly effective influencer of consumer behavior. Reviews, testimonials, and ratings only help to validate potential buyers and quite often to sway yes-or-no purchasing decisions (Kliestik et al., 2022). But in reality, there is a growing appreciation of the role user-generated content can play in bolstering brand credibility, particularly since the consumer is much more likely to believe fellow shoppers over conventional advertising. Therefore, how the social and digital have combined reflects contemporary consumer behavior in e-commerce and presents compelling reasons why retailers must alter their approach (Khodabandehlou et al., 2017).

Traditional Analytical Approaches

As per Liu et al. (2019), traditional analytical studies of consumer behavior have laid an important foundation for understanding consumer activities but are being challenged these days by how complex and heavy the data coming out from consumers are. Most of the existing techniques include regression, cluster analysis, and cohort analyses, which derive insights from the preferences and behavior choices of consumers through the use of structured data: predefined variables lead to a probability of predicting outputs. However, their applicability is often limited in the face of large-scale, dynamic data environments typical of contemporary e-commerce platforms.

According to Luo et al. (2022). one of the major limitations of traditional methods is their inability to handle the unstructured data that dominates digital interaction, brought about by social media, customer reviews, and more. In such approaches, traditional analytics fails to amalgamate various sources of data, hence yielding fragmented insights that may not capture all the subtleties of the behavior of a consumer. More than this, these approaches lack flexibility because they are typically rigid and therefore not able to change with speed when consumer preferences or market conditions change. For that reason, retailers who just stick to conventional analytical approaches are likely to find themselves at a disadvantage in this highly competitive market.

Moreover, traditional approaches demand a sizeable amount of manual input and hence expertise, causing them to fall behind in a scalable manner. There is a stark need for improved analytical tools when the electronic commerce business experiences growth and enormous volumes of raw data are surfacing. Hitherto impossible to achieve through traditions, these machine learning techniques have been adopted primarily to process any large volume of raw data in just real-time discoveries of patterns uncovering insights where humanly very impossible. This is a paradigm shift in the understanding of consumer behavior, allowing retailers to proactively respond to market dynamics as they transition from traditional analytics to machine learning (Ma & Sun, 2020).

Applications of Machine Learning in E-commerce

As per Syam & Sharma (2018), advanced models developed on consumer segmentation, recommendation systems, and demand forecasting have taken machine learning to a whole new course of the development curve for e-retail. As mentioned, strong suits of the machine learning algorithms include data analysis of larger magnitudes, extraction of useful patterns from those, and predictions of events that could result from prior activities. Examples of customer segmentation could include clustering algorithms which categorize consumers into clusters depending on purchasing habits, preferences, and demographic features. This element also enables retailers to concentrate their marketing efforts and product offerings on key segments, enhancing the quality of the shopping experience and improving conversion rates.

Retrospectively, the most salient use of machine learning in e-commerce is related to recommendation systems driven by both collaborative and content-based filtering. These recommend products consonant with personal taste by analyzing what the consumer clicks on or views to drive sales considerably. Scantly, complex sets of algorithms are utilized by Amazon for recommendation and Netflix, increasing user engagement by manifold. Plenty of success stories say e-commerce giants claimed huge revenue increases due to the right recommendation systems in place (Zhou et al., 2021). . Using machine learning will give way to a shopping journey that becomes very personalized. If so, loyal consumers and more repeat purchases are likely. Another very important area where the inroads of machine learning have been considerably felt is in demand forecasting.

The machine learning models, based on historical sales data, seasonality, and other external factors, predict future demand with uncanny accuracy and enable the retailer to optimize inventory management by reducing the risk of stockout or overstock situations. Improved demand forecasting leads not only to better operation efficiency but also to better customer satisfaction, since the probability a consumer will find in stock what he or she wants is much higher. This application of machine learning in demand forecasting justifies the belief that more insightful decisions with a strategic outlook are derived through data analysis and insights for enhanced profitability (Yoganarasimhan, 2020).

Gaps in the Research

Despite the promising applications of machine learning in e-commerce, many research gaps persist, especially concerning the unique dynamics characterizing the U.S. market. Although several machine learning techniques have been extensively studied, tailored approaches that adequately address the specific characteristics and behaviors of U.S. consumers are needed. This means that the diversity of the American market, which is influenced by cultural, economic, and regional factors, demands a more profound understanding of consumer behavior than what generic models can provide. Therefore, future research should focus on developing machine learning frameworks that consider such unique market dynamics, thus offering retailers more relevant and actionable insights.

Furthermore, there is a great lack of comparison studies regarding the results obtained by different machine learning methods in developing optimized online selling strategies. Although several papers show the efficacy of some algorithms, an overall analysis that would show the performance of various methods within different real e-commerce environments is needed. In addition, this study would not only confirm which different approaches perform but also inform a retailer of what tool will most suit the requirement. Knowing which machine learning techniques work best in which contexts can help retailers make informed decisions to improve their online sales strategies.

III. Data Collection and Exploration

Dataset Overview

The consumer behavior dataset contained complete data on various aspects of the customer's behavior in online retail. The dataset consisted of transaction records for customer purchase history, items purchased, frequency of purchases, and values of transactions. It also contained browsing history data that would point out the patterns of user interaction, such as pages visited, time spent on each page, and views of different products to draw fine-grained inferences on consumer interest and preference. Demographic information on age, gender, location, and income level enhanced the dataset further and, therefore, was useful in segmentation and targeting. The data was collected from a wide array of e-commerce websites that provided transactional and behavioral data with social media insights to consider consumer sentiment and trends. Feedback surveys provided qualitative data through direct opinions and experiences of consumers, thereby making the dataset rich for analysis.

Data Preprocessing

The code snippet that was formulated represented some of the data preprocessing steps to prepare the dataset for analysis. The first step entailed filling in missing values, dropping rows where there are missing values in critical columns, and filling in remaining missing values with appropriate strategies: 0 for numeric columns, and "Unknown" for categorical columns. The second step entailed converting columns to their correct data types, such as date and integer. The third step comprised feature engineering which was applied by adding a new feature in the dataset-"Order Month-Year" and "Customer Loyalty Duration." The fourth stage encompassed cleaning of invalid or outlier values by removing rows with negative or zero values in critical columns and unrealistic ages; numeric columns are normalized by applying Min-Max-Scaler, and encoding categorical variables is conducted using Label-Encoder. Lastly, it removes duplicated rows.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis in itself is the first and imperative step for almost any work, involving data in the Data Science domain. Visualization and statistical techniques of summarization and descriptions were involved in EDA. Using the EDA chart enabled the

analyst to get an estimation of the distributions of the given data, identification of hidden patterns, and detections of outliers that disclose hidden bonds between variables. These plots let one find the occurrence of a pattern, anomalies, and problems he might not perceive by just analyzing raw numbers. EDA is not about creating pretty graphs but rather about asking questions, formulating hypotheses, and guiding further steps of analysis and modeling. It eventually helped us understand the data better and hence make better decisions or analyze data more effectively.

Sales by Region

The formulated code snippet was executed to create a visualization of the sales data based on the region in Python. It first grouped the data based on the 'Region' column, summing up the 'total' sales of each region. Then, a seaborn bar plot was generated from this aggregated data. The plot was further customized with appropriate titles and x and y labels besides a color palette to make it more readable. Lastly, plt.show() was used to display the created bar chart representing sales performance over varied regions as displayed below:

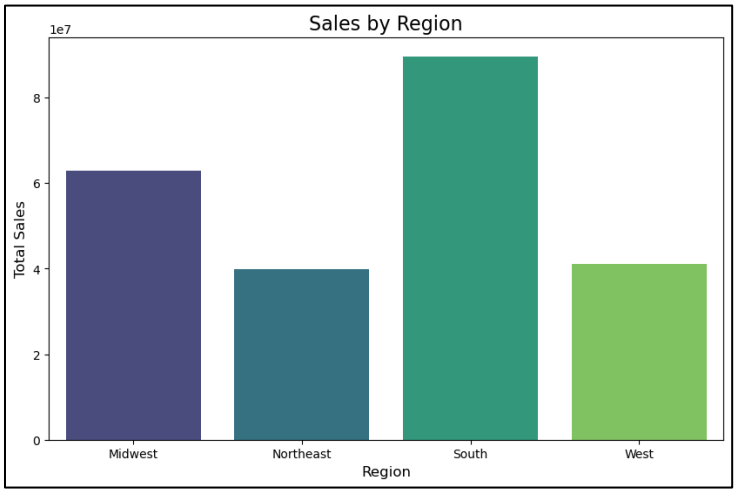

Figure 1: Displays Sales by Region

The "Sales by Region" bar chart above shows the total sales by each region. The South region leads in sales volume, at about 9 million, followed by the Midwest region, which is at about 6 million. The West region had sales of approximately 4 million, while the Northeast region recorded the lowest sales volume at approximately 4 million. This chart shows a great gap between sales in the South and sales in the rest of the country. Therefore, this may well point to the South being a priority market for the firm.

Order Density by Day and Hour

The implemented Python snippet was meant to check order density by day of the week and hour of the day, meaning it extracted first the day of the week and then the hour from the column 'order_date'. After that, the data was grouped by 'day_of_week' and 'hour', and then the counts of orders were made for every group. This aggregated data was then visualized as a heatmap using seaborn with the coolwarm colormap. Then, it puts a title to it, labels the x and y axes, and applies a tight layout. Finally, it displayed the created heatmap with plt.show ().

Output:

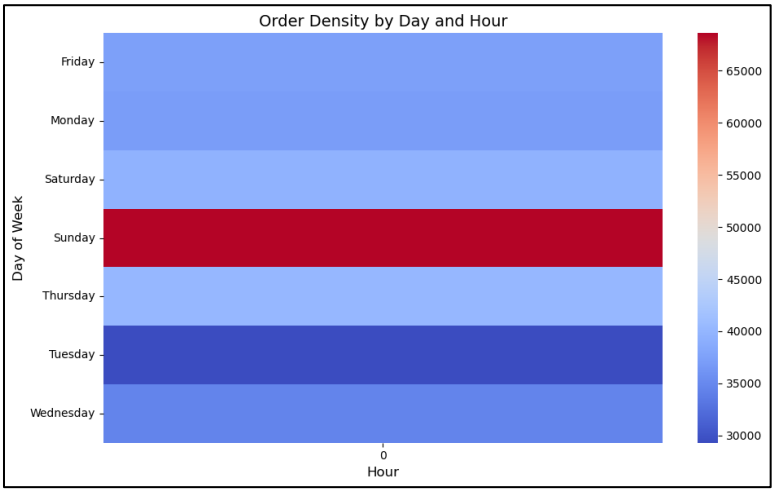

Figure 2: Displays Order Density by day and Hour

The "Order Density by Day and Hour" heatmap visualizes the distribution of orders across days of the week and hours of the day. The color intensity is according to the number of orders placed; the darker the color, the higher the order density. Therefore, Sunday seems to have the highest order density, followed by Friday, Monday, and Tuesday. Wednesday and Thursday have relatively low order densities. It has indeed indicated the trend wherein this business peaks during the weekend and at the start of the week. Further investigation can provide further detailed information about what specific hours these days are experiencing high orders, which can be beneficial in optimizing staffing and inventory.

Top 10 Customers by Revenue

The code snippet in Python visualized the revenue contributions of the top 10 customers using a radar chart. First, it identifies the top 10 customers by their total revenue using the group by and largest functions. Then, it prepares the data for the radar chart by creating a list of customer IDs as categories and a list of their corresponding revenue values. This then computes the angles for the radar plot and further normalizes the values with angles to obtain a closed polygon. It now creates a polar subplot using matplotlib and plots the radar chart, with an additional title and labels. In this way, it will have a very clear view of what revenue comes from each of those top customers to more easily compare the data and take note of exactly who the more valuable clients will be.

Output:

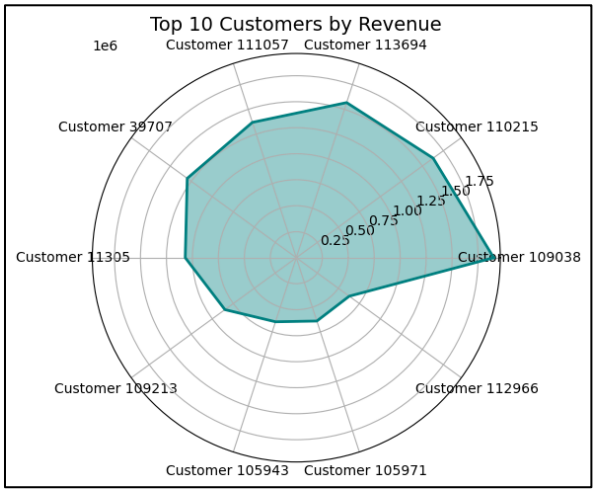

Figure 3: Visualizes Top 10 Customers by Revenue

This chart is a polar bar chart of the top 10 revenue customers to show their comparative share of total sales. Each axis represents one customer, and the radial distance corresponding to that axis reflects the revenue generated by that customer. More specifically,

Customer 110215 has by far the highest revenue, way over any of its peers, while Customers 109038 and 111067 are the next most highly ranked, albeit at lower levels. Customers like 38707 and 112966 form a middle contribution, though. This could provide great avenues for targeted marketing or engagement. Given the spread of revenue from these customer segments, it indicates there is heavy reliance on just a few key customers, and thus reinforcement of these ties becomes very vital to improve lifetime value coming from them. This graph visually summarizes revenue dynamics among top customers and serves to aid strategic decisions regarding customer relationship management.

Customer Journey: Order Over Time

The executed Python code snippet computed a line plot to visualize the order history of individual customers over time. It first began with converting 'order-date' into date-time format followed by grouping this column along the 'Customer-ID', counting the number of orders placed by one customer on certain dates. This code looped over the top 10 customers and plots their order history as lines in the same graph. The plot is further enhanced by adding a title, labels for both the x and y axes, a legend, and grid lines in a way that best conveys the readability of the data. This kind of visualization enables the visual comparison of order patterns across different customers and can be used to identify trends, seasonality, and customer behavior over time.

Output:

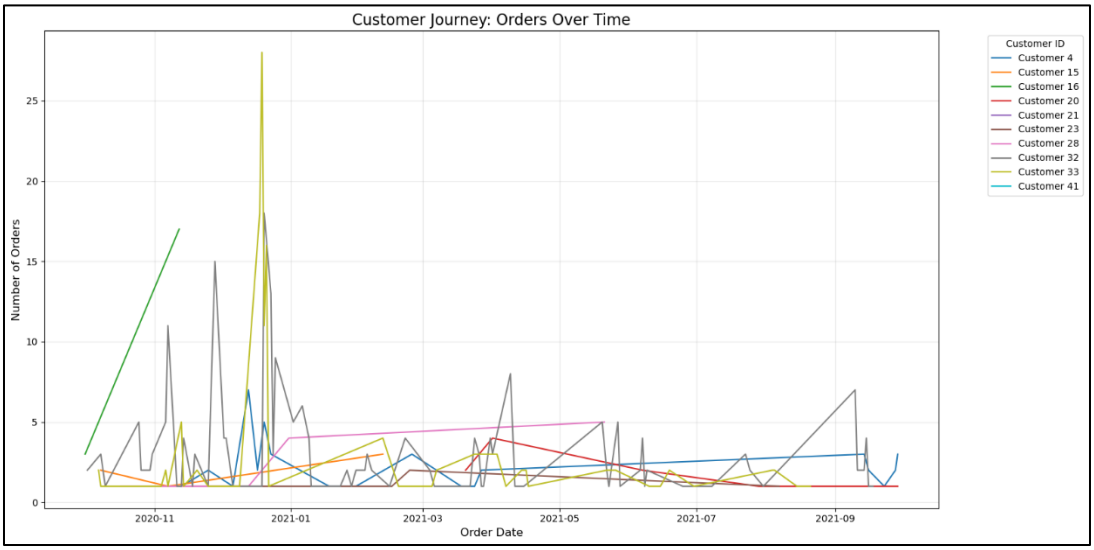

Figure 4: Exhibits Customer Journey: Orders Over Time

The graph showing the customer journey over time is enlightening, as it provides a proper view of the order frequencies of different customers between late 2020 and September 2021. Moreover, Customer 4 shows two striking spikes in order volume, the first being in late 2020, indicating a probable promotional period or high level of engagement in the year just gone. In contrast, Customer 15 and Customer 16 have demonstrated consistent ordering patterns throughout the period under observation, with relatively stable but lower order counts, which characterizes loyal but less frequent purchasing behaviors. The erratic trends in Customers 21 and 28 indicate fluctuating engagement that may require further investigation to understand the root causes, whether related to satisfaction levels, market conditions, or promotional effectiveness. The general trend shows different levels of engagement from customers; some customers are very loyal, whereas others have been purchasing inconsistently, meaning marketing strategies will need to be personalized to retain more customers and ultimately drive higher order volumes.

Revenue by Category and Year

The Python code snippet was computed to display a stacked bar chart of the contribution of various categories to revenues over time. It first grouped data by category and year, then calculated the total revenue of each category in each year. Then it pivoted the data to create a data frame ready for plotting with years as an index and categories as columns. Next, it produced from the pivoted table a stacked bar chart, whereby each bar provides the total revenues in one year while, within this, each different-colored part of this is what constitutes that total for every one of these different categories. The plot generated a tailored title, labels around the x and y-axis, and provided a colormap displaying a legend box. The following chart succinctly outlines how revenues have changed for each category over time, so that comparisons may be easily made, trends identified, and the shifting of the market.

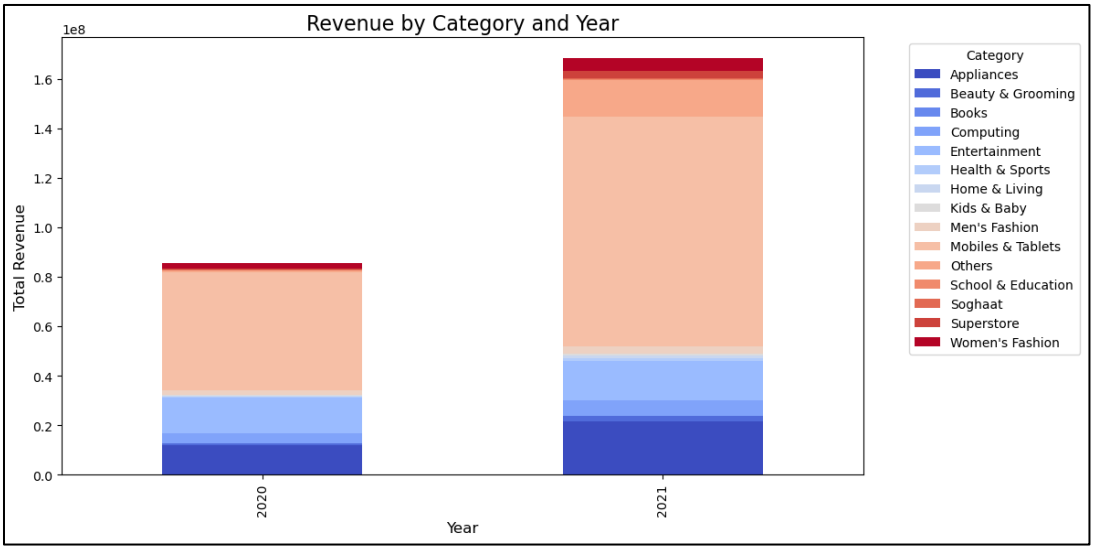

Figure 5: Portrays Revenue by Category and Year

This Python code snippet analyzed sales data by age group. First, it defined bins and labels for placing customers into an age group. Then, it created a new column in the data frame called 'Age Group' using the defined bins on the column 'Age'. It then grouped the data by 'Age Group' and calculated the total sales in each group. Eventually, a seaborn plot in Python was used that created a bar visualization of the number of sales attributed to each possible unique demographic using age. Therefore, it then provided a chart labeling the title with meaningful x and y labels, and coloring bar graphics with pastels for readability: This analysis explained the performance and built actionable implications for targeted market strategies based on the age demos.

Output:

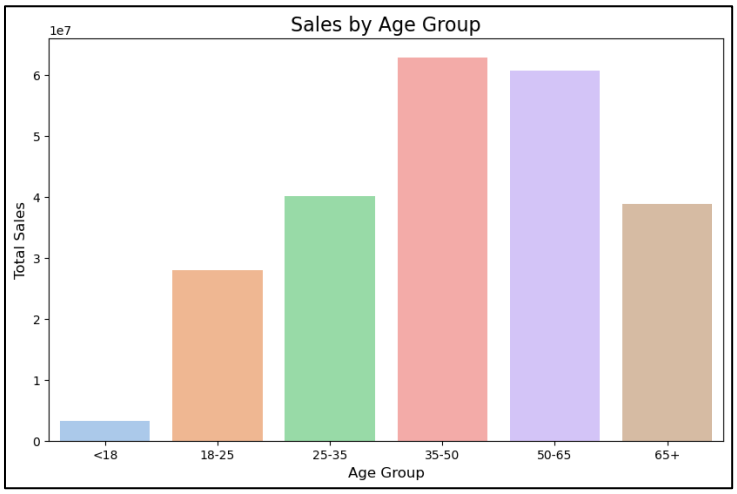

Figure 6: Showcases Sales by Age Group

The chart on sales by age group shows very interesting purchasing trends across the different demographics. The age bracket of 35-50 years is the highest contributor to total sales, with approximately $6 million in sales, which could mean that this demographic is more financially stable or has a stronger inclination to purchase the products being offered. The next best group is the 50-65 age bracket, at approximately $5 million, indicating that there could be significant market potential in older consumers who are more appreciative of the quality and reliability of the products. The 25-35-year-old bracket generates around $3 million, representing a younger demographic that may appreciate other aspects of the product or different pricing. At the low end, the <18 age bracket records very few sales, which would indicate that younger consumers have not yet commanded much purchasing power or brand loyalty. Overall, these statistics underline the need to zero in on effective marketing across age groups to capitalize on the most lucrative segments while considering ways of engaging younger consumers as they mature.

Sales by Gender

The computed code in Python interpreted the sales data by gender. First, it grouped the data using the 'Gender' column and computed the total sales for each gender. Further, aggregated data is used to plot a bar plot using the Seaborn library. It was customized with a title, x, and y labels, and a color palette for better readability. Finally, the plot.show() command generated the bar chart to show the above for sales performance across different genders.

Output:

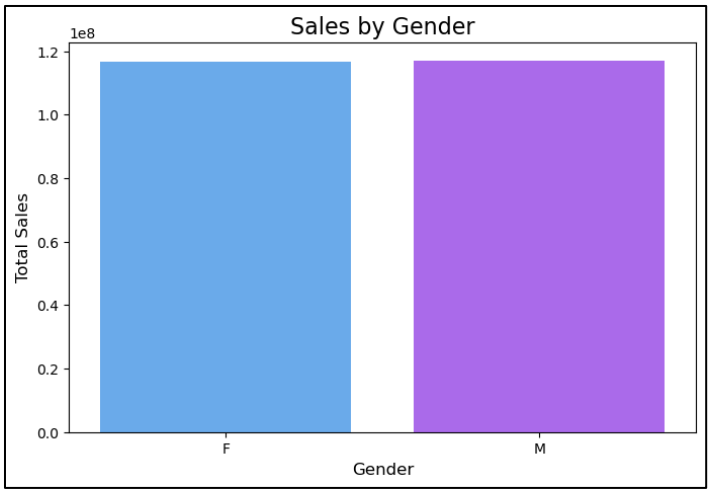

Figure 7: Illustrates Sales by Gender

The chart above on sales by gender is relatively well balanced between female and male consumers, with total sales at approximately $1.1 million across both genders. Sales for females (F) are slightly higher than for males (M), indicating that female customers may be more involved or attentive to the products offered. This would imply that there is some parity in sales, and therefore marketing strategies should be all-inclusive and appeal to both genders since neither group dominates the purchasing landscape. The data reinforces the notion that businesses can benefit from understanding the preferences and behaviors of both male and female consumers to optimize product offerings and marketing campaigns, potentially enhancing overall sales performance.

Sales by Category and Region

The provided Python code snippet generated a stacked bar chart to visualize sales by category and region. It first grouped the data by 'category' and 'Region' and calculated the total sales for each category in each region. Then unstack grouped data into the Data Frame was sketched out for the plotting, keeping 'categories' at the index and ' regions' at the columns. It plotted the bar chart stacked, from a pivot table of total sales against the respective categories and represents different colored segments inside each bar corresponding to each region's sales contribution. The plot configured title, x-label, and y-label with labels for legend, and colormaps, respectively, for the plotting of results. The bar chart below represents sales distribution in various regions for each category, thereby making it easier to compare and identify regional sales trends and variations.

Output:

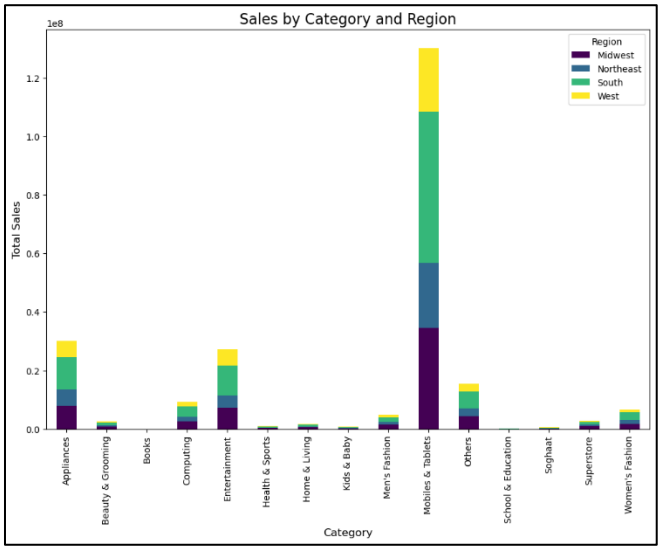

Figure 8: Depicts Sales by Category and Region

The sales chart by category and region indicates some rather interesting trends in the categories of products and the geographical territories where the sales were made. "Men's Fashion" and "Mobiles & Tablets" top the graphs in total sales, while the Northeast and Midwest regions are most dominating, hence showing that technology and apparel have a very huge market in these areas. Beauty & Economy and Health & Sports sectors, though appear relatively low compared with others, therefore they may have needed proper marketing campaigns to arouse consumer interest in these segments. The sales in the South region appear to have been distributed diversely in various sectors and the Western region seems to be more distributed uniformly across sectors about wide and diversified consumers' interests. The overall implication of this data is that regional marketing strategies should be different for different product categories, and thus businesses can accordingly plan their strategies based on regional consumer behavior.

IV. Machine Learning Methodology

Model Development

The task of understanding customer behavior in the USA, for a better online sales strategy, is quite broad. Machine learning models help in the analysis of complex data sets through which a business can gain actionable insights. The analyst implemented accredited and credible models, such as Random Forests, Logistic Regression, and Gradient Boosting Classifiers, that are useful in various dataset analyses related to customer behavior. Each of these methods has its strengths and can be used judiciously based on the nature of the data about customer behavior.

Random Forest is an ensemble learning technique that functions via the curation of multiple decision trees and outputs the mode of their predictions for classification tasks or the mean prediction for regression tasks. Random Forests can easily deal with big data in high dimensions; hence it effectively capture the complicated patterns within the data about consumer behavior. With the averaging mechanism, the overfitting robustness characteristic also enhances the predictive accuracy of Random Forest, which makes it an efficient choice for models where consumer profile and behavior would vary.

Logistic Regression is a statistical method designed for classifications of the binary target and used to develop an understanding of the relationship between a dependent binary target variable and one or more independent variables. Its strongest advantages involve interpretability in ways that the stakeholders can fathom the effects of different influences on consumer decisions. In particular, it is useful in any research objective directed toward identifying major predictors of customers' behavior, such as the likelihood to buy, which may be induced by demographic characteristics or past purchasing history.

Gradient Boosting Classifiers is another powerful one, that does well when the predictive accuracy of a model is of prime importance. It works by constructing trees one after another, with each tree correcting the errors of the previously built tree. It is very flexible and can be tuned to minimize a wide range of loss functions, which makes it suitable for capturing complex nonlinear relationships in customer behavior data. It handles missing values and also has regularization techniques.

The choice of these models is supported by the characteristics of the dataset and the overall objectives of the research. As the customer behavior data are complex and highly dimensional, models such as Random Forest and Gradient Boosting offer the opportunity to capture non-linear interactions and relationships. However, Logistic Regression presents a relatively easy way of interpreting the influence of specific variables, which can then be used for guiding strategic decisions because of interpretability.

V. Results and Insights

Model Performance Analysis

1) Gradient Boosting Classifiers Modelling

The Python code for a Gradient Boosting Classifier model instantiated a Gradient-Boosting-Classifier object with a random state of 42 for reproducibility. Then, it fitted the model on the training data X-train, and y-train using the fit () method. Finally, the trained model predicted the labels for the test data X_test and stored these predicted labels in y_pred_gb. Finally, the code printed the accuracy score and the classification report, which included detailed metrics, like precision, recall, and F1-score for each class to gauge the performance of the Gradient Boosting Classifier model.

Output:

Table 1: Gradient Boosting Classifier Report

Gradient Boosting Classifier Results:
Accuracy: 0.4738069698405967
Classification Report:

	precision	recall	f1-score	support
0	0.55	0.60	0.57	30427
1	0.13	0.06	0.08	54
2	0.17	0.00	0.01	857
3	0.42	0.63	0.51	25551
4	0.00	0.00	0.00	8
5	0.60	0.02	0.04	7268
6	0.17	0.04	0.06	216
7	0.11	0.05	0.07	21
8	0.00	0.00	0.00	13
9	0.00	0.00	0.00	2
10	0.00	0.00	0.00	9
11	0.41	0.23	0.29	14680
12	0.38	0.01	0.01	1068
accuracy			0.47	80174
macro avg	0.23	0.13	0.13	80174
weighted avg	0.48	0.47	0.44	80174

The table above shows results from the Gradient Boosting Classifier. The confusion matrix indicates the overall accuracy was about 47.4%, with 80,174 total instances. Precision and recall for every class revealed pretty good performances; specifically, class 0 showed a precision of 0.55 and a recall of 0.86, which indicates high True positives that represent this class against misclassifying a big instance number. While class 1 has a much lower precision of -0.17 and a recall of only 0.06, this further shows the complexity of correctly predicting this particular class. This is tough to face from another perspective, too, concerning both accuracy and class weight, as by calculating their f1 score, class 1 obtains the minimum value reaching f1-sc=0.10. The average overall for the f1-score weighted makes out to be 0.48 testifying to weak performance on behalf of this model across subjects.

2) Random Forest Modelling

The Python code snippet implemented a Random Forest Classifier model. First, it instantiated a Random-Forest-Classifier with 100 decision trees and set the random state to 42 for reproducibility. Then, using the fit() method, the model was fitted on the training data, X_train and y_train. Thereafter, the fitted model was used in predicting the labels for the test data, X_test, and the predicted labels are stored in y_pred_rf. Finally, it printed out the accuracy score and the classification report that includes the detailed metrics: precision, recall, and F1-score of each class to gauge the performance of the Random Forest Classifier model.

Output:

Table 2: Random Forest Classifier Report

Random Forest Classifier Results:
Accuracy: 0.5017711477536358
Classification Report:

	precision	recall	f1-score	support
0	0.57	0.64	0.60	30427
1	0.27	0.17	0.21	54
2	0.13	0.07	0.09	857
3	0.51	0.55	0.53	25551
4	0.29	0.25	0.27	8
5	0.29	0.20	0.24	7268
6	0.22	0.09	0.13	216
7	0.20	0.05	0.08	21
8	0.14	0.08	0.10	13
9	0.00	0.00	0.00	2
10	0.00	0.00	0.00	9
11	0.41	0.35	0.38	14680
12	0.13	0.07	0.09	1068
accuracy			0.50	80174
macro avg	0.24	0.19	0.21	80174
weighted avg	0.49	0.50	0.49	80174

The table below shows the results of a Random Forest Classifier for an overall accuracy of about 50.2% at 80,174 instances. There is a huge variation of precision and recall across classes; class 0 has the highest precision, 0.57, and recall, 0.67, showing reasonable ability to identify instances correctly. Whereas other classes like class 1 and class 2 also provide very poor performances with class 1 obtaining as low as 0.13 in terms of precision, the class only provides an accuracy recall value of 0.08. In both these measures, especially where support is lower, this depicts some areas such that if all are summed, then weighted f1 results will amount to an average measure of 0.50. Overall, while representing the same subject area, there are issues such that classification accuracy will suffer because classes show features in the set that could fit other categories quite adequately.

3) Logistic Regression

This code snippet in Python implemented the logistic regression model for classification. It first selected the features that this model would require; these are features like 'qty_ordered', 'price', 'value', etc., to variable X. It then trained the data using the train_test_split. In this section, a Logistic Regression model was instantiated and then fitted with the training data through a call to the fit() method; after that, it will be used to make predictions on the test set. Finally, it printed out the accuracy score and classification report, including the detailed metrics for precision, recall, and F1-score in each class to evaluate the Logistic Regression model performance.

Output:

Table 3: Logistic Regression Results

Logistic Regression Results:
Accuracy: 0.37540848654177167
Classification Report:

	precision	recall	f1-score	support
0	0.38	0.85	0.52	30427
1	0.00	0.00	0.00	54
2	0.00	0.00	0.00	857
3	0.38	0.16	0.22	25551
4	0.00	0.00	0.00	8
5	0.00	0.00	0.00	7268
6	0.00	0.00	0.00	216
7	0.00	0.00	0.00	21
8	0.00	0.00	0.00	13
9	0.00	0.00	0.00	2
10	0.00	0.00	0.00	9
11	0.36	0.00	0.01	14680
12	0.00	0.00	0.00	1068
accuracy			0.38	80174
macro avg	0.09	0.08	0.06	80174
weighted avg	0.33	0.38	0.27	80174

The table above exhibits the report of a Logistic Regression model, indicating an overall accuracy of approximately 37.5% across 80,174 instances. The precision and recall vary greatly across the classes. For instance, class 0 had high precision with 0.85 and low recall at 0.52; while good for identifying true positives, it is very poor regarding catching actual instances. Other classes, in particular class 1 and class 3, are very bad, with the f1-scores of 0.16 and 0.22, correspondingly, which shows that a model had a hard time making correct predictions on these classes. The weighted average F1 score reaches 0.38 and describes the overall modest success of the model so far, requiring further tuning if class differentiation needs to be performed.

Consumer Behavior Patterns

Comprehending consumer behavior trends is paramount for efficiently steering purchases and enhancing customer satisfaction. The main insights reveal that price sensitivity is one of the most important factors influencing consumer choice. Many are on the lookout for discounts or promotions, especially in highly competitive markets where choices are many. Reviews and ratings regarding a product are also significant, as people rely on others' experiences to estimate the quality and reliability of the product. Positive reviews can help build trust and drive sales, while negative feedback can prevent buyers. Besides, social media has given a full swing to reviews as people now use Instagram and Facebook to discover and then validate products before making a purchase.

Customer segmentation analysis also brings out the diverse shopping preferences among different consumer groups. For example, millennials may be more concerned with the sustainability and ethics of the brand, while older generations might be more concerned with the quality of the product and value for money. Segmentation by income levels also shows that different levels of price sensitivity exist; for higher-income consumers, the price may be less of a driver, with exclusivity or brand prestige more important. This nuance, if understood, helps the businesses in customizing their marketing approaches, so that the same are related to the needs and preferences of each consumer segment.

Actionable Sales Strategies

To capitalize on insights into consumer behavior, organizations should consider deploying personalized promotions that cater to individual shopping preferences. Data analytics may be used to gain insight into particular consumer behaviors and to construct targeted marketing campaigns featuring relevant products and offers. Examples include personalization of discounts on products that customers have bought before to improve customer loyalty and increase conversion rates. Bundling products together at a

discounted price can also incentivize consumers to buy more, particularly when the bundled items complement each other, appealing to the desire for convenience and value.

Another important strategy will be to enhance navigation and improve the user experience of the website. A friendly interface will go a long way in enhancing the user's shopping experience and reducing cart abandonment. Making checkout more seamless, showing clear categorization, and adding intuitive search can help a customer locate just what they may be looking for. Additionally, customer reviews and ratings up front on products will help engender trust with people so they can feel more confident about the likelihood of making good purchases. Therefore, mobile responsiveness is important to consider, too, because more and more consumers shop this way on their smartphones. Focusing on these elements will help businesses create a more engaging and satisfying online shopping environment, leading to higher sales and customer retention.

VI. Practical Applications

Implementing Machine Learning in E-Commerce

Some critical steps are involved in integrating machine learning models into online retail platforms to make them function seamlessly and enhance customer experiences. First, the business has to define its objectives, which could be to enhance customer recommendations or optimize pricing strategies. Second, data collection is vital; an e-commerce platform should amass extensive datasets on customer behavior, transaction history, and product information. The data will serve as the backbone for training the machine learning models. Once the data is collected, the next step will be to select appropriate algorithms based on the specific objectives. The common algorithms include collaborative filtering for recommendations and regression models for dynamic pricing. The models should subsequently be trained and then tested for their accuracy and effectiveness before deployment.

Some of the key examples of machine learning applications in e-commerce are real-time recommendations and dynamic pricing mechanisms. For instance, a recommendation system could make suggestions based on the history of a user's browsing and purchases, recommending products that best match their preferences to increase the likelihood of more purchases. A good example could be how Amazon has "Customers who bought this item also bought," which efficiently guides the users to complement their products. Dynamic pricing, on the other hand, relies on machine learning algorithms that adjust prices based on demand fluctuations, competitor pricing, and customer behavior. This strategy allows retailers to maximize revenue by ensuring prices remain competitive while also appealing to price-sensitive customers.

Improving Customer Retention

The backbone of e-commerce business growth is to enhance customer retention strategies, whereby loyalty and repeat purchases are significantly enhanced. Loyalty programs are a core means of retention, offering some sort of incentive for customers to return. These could be in the form of points that are redeemable or tiered membership providing exclusive benefits. Other key strategies will include personalization: where customer data is used to craft very specific marketing messages and product recommendations. For example, an email featuring products related to what a customer has bought in the past will increase engagement and conversion.

Engagement also plays a very important role in retention. The companies should use all possible channels of communication: social media, e-mail, and SMS, to notify customers about the current promotions, new products, and news about the company. Interactive content such as quizzes or polls will create a community feeling and make customers feel important. Also, feedback from customers and acting upon it is a clear indication that the brand listens to its audience's needs, which strengthens loyalty more and more. By implementing these strategies, e-commerce companies can build meaningful relationships with their customers and, therefore, enjoy better retention rates and lifetime value.

Scalability Across Industries

This mechanism of the implementation of machine learning to improve customer retention is well-adaptable in many e-commerce industries, ranging from fashion and electronics to groceries. In the fashion industry, for example, machine learning can be applied to analyze the trends and forecast consumer preferences to enable the retailer to stock items that will sell well. Recommendation engines could suggest outfits or create lookbooks, making the shopping experience even more engaging with curated options. In electronics, machine learning can optimize inventory by predicting what products will be in demand during which season, based on seasonal trends and historical sales data.

Other methodologies discussed would also enable grocery e-commerce to create a more personalized touch toward the experience of consumers when doing their grocery shopping. Using machine learning, a grocery retailer can make recipe suggestions to customers based on their purchase history and dietary preferences. Loyalty programs can include features like personalized discounts on frequently purchased items or seasonal product deals. E-commerce companies, therefore, can use

machine learning and retention strategies fitted to the peculiar demands and nature of each industry, which will lead to better business growth and agility in a dynamically changing marketplace.

VII. Discussion and Future Directions

Impact on the U.S. E-Commerce Market

Machine learning can revolutionize the way online retailing is approached in the U.S., as it has the potential to make full use of consumer data on a large scale for more nuanced decision-making. While integrating machine learning algorithms, retailers can develop highly personalized shopping experiences that best meet the preferences and behaviors of individual customers. For example, predictive analytics can forecast buying patterns in the future, thus enabling retailers to create an optimal inventory and marketing campaign. Machine learning can also enhance customer service by providing real-time responses through chatbots and other automated systems, reducing response times and enhancing overall customer satisfaction. This move toward data-driven strategies not only enhances operational efficiency but also positions retailers to better meet the evolving demands of consumers in an increasingly competitive landscape.

These changes have broader implications for the wider diffusion of customer-centric business models across the industry. With more accessible machine learning tools, businesses can be better equipped to pay greater attention to customer contact points. The transition in companies fosters a continuous improvement in business functions, owing to continuous feedback loops starting with customer interaction to product development and service improvement. Firms holding such customer-centric practices tend to have brand loyalty, thus fostering a long-term relationship with the customer base. This would be continued growth in the e-commerce sector.

Limitations of the Study

Despite the promising insights into the impact of machine learning on e-commerce, notable limitations do exist in this study. One challenge relates to data quality. As it is with all machine learning algorithms, they are only as good as the data they have been trained on. Poor datasets lead to poor results, and the generalization of such findings is limited. Moreover, the fast-changing nature of consumer behaviors makes their capture and analysis quite difficult. With external factors such as economic conditions, changes in technology, and cultural shifts continuing to make preferences change, it may be difficult for static models to keep pace with such changes, yielding conclusions that are outdated and less effective strategies.

Again, besides these two broad segments of the industry, it could be incomplete to focus merely on e-commerce when a lot of diversities are noticed among various consuming behaviors within each industry. Such a limitation gives a serious setback in the case of applying such findings because every varying market dynamic or expectation by the consumer would already determine the mode of application of the machine learning strategies. Thus, any further research study shall attempt to eliminate the shortcomings using rigorous methodology and controlling factors associated with data quality and fluid consumer behavior.

Future Research Opportunities

Numerous opportunities for future research can further enhance our understanding of machine learning applications in e-commerce. Further areas of exploration may involve more data sources, including social media interactions and Internet-of-Things devices. Integrating these diverse datasets can provide richer insights into consumer preferences and behaviors, enabling more nuanced analyses and tailored marketing strategies. For instance, sentiment analysis in social media could be used to determine precisely how the general public perceives their products, thus enabling retailers to make related adjustments.

Moreover, advanced model integration, such as reinforcement learning, allows great scope for strategy development in an always-changing dynamic to reach consumers. Unlike the traditional supervised learning models, reinforcement learning allows systems to learn through interaction with the environment, and hence, it is especially suitable for the real-time optimization of pricing strategies, inventory management, and personalized recommendations. E-commerce businesses can use these advanced techniques to gain a competitive edge, respond promptly to market dynamics, and, therefore, serve their customers even better. In other words, any research in these areas will strengthen not only the understanding of the impact of machine learning on e-commerce but also contribute to the development of far more effective and adaptive retail strategies.

Comparison of All Model

Suitable code snippet compared the performance of three classification models: Logistic Regression, Random Forest, and Gradient Boosting. Further, it calculated the accuracy score for each model using the function accuracy_score and stored it inside a dictionary; then, it computed which one has the highest accuracy with max. Finally, it printed out a comparison table regarding the accuracy of all the models with the best performers. This code helped in evaluating and selecting the best classification algorithm for the dataset in concern based on their accuracy scores.

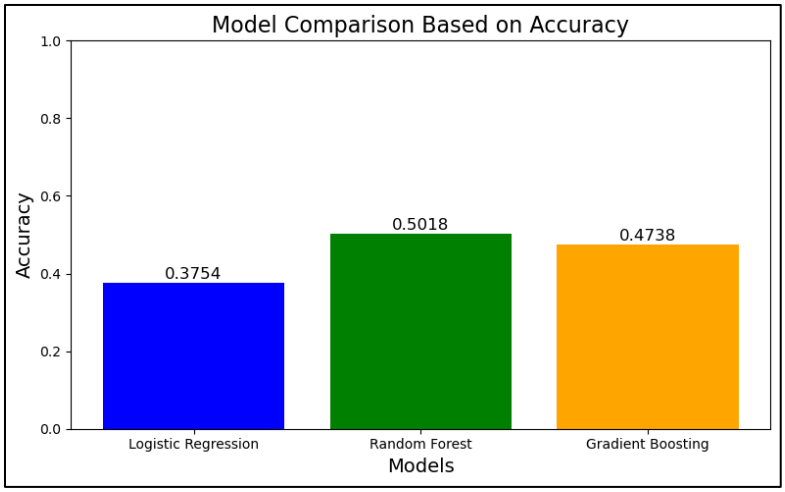

Figure 2: Displays Model Comparison Based on Accuracy

This chart compares the performance of three machine learning models: Logistic Regression, Random Forest, and Gradient Boosting, all against the performance of a particular task. Logistic Regression had an accuracy of about 37.5%, which is the lowest accuracy achieved by the evaluated models. On the other hand, Random Forest turned in a strong performance, having an accuracy of about 50.1%, reflecting that it is efficient in picking up complex patterns in the data. It is closely followed by Gradient Boosting, which has an accuracy of 47.3%, hence the potential in the various ways it might improve the performance of the predictions but does not outperform Random Forest. By that comparison, although Logistic Regression may be pretty handy in conducting easier tasks, for scenarios involving higher accuracy or sophistication, better options exist like Random Forest or Gradient Boosting. It thereby underlines the essentiality of the model selection criterion concerning data and goal characteristics.

VIII. Conclusion

This research aims to apply machine learning techniques to develop a strategic online sales strategy through deep consumer behavior analysis. This research paper focuses on a consumer behavior analysis based on U.S.-based datasets underlining American consumers' unique characteristics and preferences. The consumer behavior dataset contained complete data on various aspects of the customer's behavior in online retail. The dataset consisted of transaction records for customer purchase history, items purchased, frequency of purchases, and values of transactions. It also contained browsing history data that would point out user interaction patterns, such as visited pages, time spent on each page, and views of different products to draw fine-grained inferences on consumer interest and preference. The analyst implemented accredited and credible models, such as Random Forests, Logistic Regression, and Gradient Boosting Classifiers, that are useful in various dataset analyses related to customer behavior. Random Forest turned in a strong performance, having relatively high accuracy, reflecting that it is efficient in picking up complex patterns in the data. Machine learning can revolutionize the way online retailing is approached in the U.S., as it has the potential to make full use of consumer data on a large scale for more nuanced decision-making. While integrating machine learning algorithms, retailers can develop highly personalized shopping experiences that best meet the preferences and behaviors of individual customers.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ballestar, M. T., Grau-Carles, P., & Sainz, J. (2019). Predicting customer quality in e-commerce social networks: a machine learning approach. *Review of Managerial Science*, *13*, 589-603.
[2] Bharadiya, J. P. (2023). Machine learning and AI in business intelligence: Trends and opportunities. International Journal of Computer (IJC), 48(1), 123-134.
[3] Boppiniti, S. T. (2022). Exploring the Synergy of AI, ML, and Data Analytics in Enhancing Customer Experience and Personalization. International Machine Learning Journal and Computer Engineering, 5(5).
[4] Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. International journal of forecasting, 35(1), 170-180.
[5] Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, *149*, 113622.

[6] Choi, J. A., & Lim, K. (2020). Identifying machine learning techniques for classification of target advertising. ICT Express, 6(3), 175-180.
[7] Feldman, J., Zhang, D. J., Liu, X., & Zhang, N. (2022). Customer choice models vs. machine learning: Finding optimal product displays on Alibaba. Operations Research, 70(1), 309-328.
[8] Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T., & Potapov, D. (2020). Digital analytics: Modeling for insights and new methods. Journal of Interactive Marketing, 51(1), 26-43.
[9] Khrais, L. T. (2020). Role of artificial intelligence in shaping consumer demand in E-commerce. Future Internet, 12(12), 226.
[10] Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), 65-93.
[11] Kliestik, T., Zvarikova, K., & Lăzăroiu, G. (2022). Data-driven machine learning and neural network algorithms in the retailing environment: Consumer engagement, experience, and purchase behaviors. Economics, Management and Financial Markets, 17(1), 57-69.
[12] Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. Expert Systems with Applications, 150, 113342.
[13] Liu, X., Lee, D., & Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. Journal of Marketing Research, 56(6), 918-943.
[14] Luo, Y., Yang, Z., Liang, Y., Zhang, X., & Xiao, H. (2022). Exploring energy-saving refrigerators through online e-commerce reviews: an augmented mining model based on machine learning methods. Kybernetes, 51(9), 2768-2794.
[15] Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. International Journal of Research in Marketing, 37(3), 481-504.
[16] Sharma, A., Patel, N., & Singh, V. (2020). Leveraging Reinforcement Learning and Bayesian Optimization for Enhanced Dynamic Pricing Strategies. International Journal of AI and ML, 1(3).
[17] Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial marketing management, 69, 135-146.
[18] Yoganarasimhan, H. (2020). Search personalization using machine learning. Management Science, 66(3), 1045-1070.
[19] Zhou, M., Chen, G. H., Ferreira, P., & Smith, M. D. (2021). Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware. *Journal of Marketing Research*, *58*(6), 1079-1100.