Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Leveraging Machine Learning for Insights and Predictions in Synthetic E-Commerce Data in the USA: A Comprehensive Analysis

Md Rafiqul Islam[1], Miraz Hossain[2], Mahfuz Alam[3], Medhat Mohiuddin Khan[4], Md Masud Karim Rabbi[5], Md Fazlay Rabby[6], Kanchon Kumar Bishnu[7], Biswajit Chandra Das[8], Md Tanvir Rahman Tarafder[9]

Abstract

The primary purpose of this research was to leverage the potential of machine learning to extract meaningful information out of synthetic e-commerce data to overcome the limitations of traditional analysis. This research focused on the consumption patterns and trends of the US e-commerce market with specific knowledge of its idiosyncrasies and challenges. The synthetic e-commerce dataset comprises a comprehensive collection of simulated transactional data designed to reflect the dynamics of an online retail environment. The dataset included detailed records of customer transactions, capturing essential information such as transaction IDs, timestamps, product categories, quantities purchased, and total transaction values. Additionally, customer demographics are represented, encompassing attributes such as age, gender, location, and income levels, which facilitate deeper insights into consumer behavior and preferences. The dataset also featured product categories that range from electronics to apparel, allowing for diverse analyses of purchasing trends across different market segments. For model choice, we applied various machine learning algorithms specific to our needs of predicting sales, customer segmentation, demand forecast, and fraud detection. Random Forest, Logistic Regression, and K-Neighbors Classifier are the models selected by the analyst. For evaluating the performance of the models, we used a suite of metrics specific to the task at hand. For the case of fraud detection, the metrics included were accuracy, precision, recall, F1-score, and ROC-AUC. The heights of the bars indicated the accuracies of the models, with Random Forest being the highest, followed by KNN, and Logistic Regression being the least among the two models. The varying heights of the bars pictorially display the comparative performance of the models, with Random Forest being well ahead of the other two models. Machine learning-driven insights have transformed e-commerce business strategies by enabling data-driven strategies to improve pricing, marketing, and stock management. Using algorithms that consider historical sales patterns, rival pricing, and market conditions, companies can implement dynamic pricing strategies that change according to real-time conditions to maximize profit margins while being cost-effective. Integrating machine learning models with real-time anti-fraud systems is a key innovation in anti-fraud measures and risk management. Using complex algorithms that examine the data of transactions in real-time enables companies to detect potentially fraudulent activity while the activity is taking place.

Keywords: *E-commerce, Machine Learning, Consumer Behavior, Sales Prediction, Customer Segmentation, USA.*

Introduction

Al Montaser et al. (2025), reported that the rapid growth of the e-commerce sector in the US has revolutionized the way the retail sector functions with a strong movement away from traditional physical stores to digital stores. The change was driven by a convergence of advancements in technologies, a change of preferences among shoppers, and the extensive usage of mobile devices that have all contributed to a smooth online purchasing experience. Statistics indicate that the US e-commerce sector experienced exponential growth, with the sales of e-commerce forecast to exceed $1 trillion within the next few years. Akter et al. (2023), argued that the meteoric growth is a testament to the need to implement data-driven strategies that are built upon insights derived from interactions with the customer base. With the e-

[1] MBA in Business Analytics, International American University, Email: me.hanifislam@gmail.com, (Corresponding Author)
[2] Master of Business Administration, Westcliff University.
[3] MBA in Business Analytics. International American University.
[4] Master of Business Administration, Westcliff University.
[5] Master of Business Administration, International American University
[6] MBA- management information system, American international university
[7] MS in Computer Science, California State University Los Angeles
[8] BS in Computer Science, Los Angeles City College
[9] Master of Science in Information Technology, Westcliff University

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

commerce sector growing by the day, companies need to appreciate the need to leverage the capabilities of big data analysis to improve customer interactions, improve operational efficiencies, and generate increased sales.

Hasan et al. (2025), indicated that in this age of the digital revolution, the scale of information that is generated by online activity is staggering. Every purchase made, page viewed, and click creates a trail of information that can deliver rich insights into customer behavior and market patterns if they are correctly interpreted. However, the complexity and scale of the information have the effect of posing a daunting impediment to traditional analytical methodologies. Traditional methodologies are applied to historical information with simplistic statistical analysis that may not fully appreciate the sophistication of modern customer behavior. According to Sizan et al. (2025a), companies will then fail to correctly forecast customer preferences, uncover emerging patterns, and learn about possible deception, ultimately failing to guide their strategic decisions accordingly. In response to the challenges that are associated with the traditional approach to analysis, the introduction of the usage of machine learning methodologies into the analytical paradigm represents a possible solution. With the support of intricate algorithms that can manage and analyze big sets of information, machine learning enables companies to uncover hidden patterns, generate practical insights, and enhance their overall capabilities of making decisions.

Problem Statement

Notwithstanding the increasing trend of basing their strategies on data-driven solutions, the majority of companies are failing to appropriately leverage e-commerce data to inform their strategies. Traditional analytical infrastructure is oftentimes ill-suited to manage the intricacies of big e-commerce data. For instance, traditional methodologies can struggle to accommodate the dynamic patterns of customer behavior influenced by a myriad of drivers, such as seasonality, marketing campaigns, and socio-economic patterns. Companies can then fail to forecast customer behavior, patterns of sales, and potential fraud with a reasonable degree of accuracy (Shawon et al., 2024)

Moreover, as per Sumsuzoha et al. (2024), the pitfalls of traditional analysis go much deeper than the question of being able to forecast with greater precision. Many companies lack the infrastructure to collect, manage, and analyze the significant amounts of information that are generated by e-commerce activity. It can lead to information silos that are rich with information that is inaccessible to the organization to deliver meaningful insights. The inability to adequately analyze e-commerce information not only hampers operational effectiveness but also the potential of personalized marketing campaigns and customer interaction strategies.

Given these challenges, the imperative is to have the organizations consider innovative means of analyzing the information. Incorporation of the methodologies of machine learning into e-commerce analysis is a potential solution that can enable companies to shatter the shackles of conventional analysis. With the aid of machine learning algorithms, companies can enhance their capabilities to analyze complex sets of information, discover patterns, and deliver a better prediction of customer behavior and sales performance (Rahman et al., 2024).

Research Objective

The primary purpose of this research is to leverage the potential of machine learning to extract meaningful information out of synthetic e-commerce data to overcome the limitations of traditional analysis. It is to develop models that can effectively perform customer segmentation, forecast demand, and identify fraud to provide the company with the means to enhance the company's decision-making capabilities. To achieve this purpose, the research will begin by generating synthetic e-commerce data that reflects the details and dynamics of the U.S. market with accuracy. The dataset will then serve as the foundation upon which to deploy a suite of machine-learning algorithms to carry out a detailed analysis of customer behavior and patterns. With clustering to segment the customer base, time series to predict demand, and anomaly detection to discover signs of fraud, the research will shine a light on the ways that machine learning can translate raw information into meaningful insights. Furthermore, the research will compare the

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

performance of the algorithms to measure the effectiveness of the different models of machine learning, both concerning their predicting value and usability. In this sense, the research will seek the best approaches to studying e-commerce information to deliver significant results to companies.

Scope and Relevance

This research will focus on the consumption patterns and trends of the US e-commerce market with specific knowledge of its idiosyncrasies and challenges. In analyzing the intricacies of the American consumption environment, the research will aim to provide information that can inform business strategies and enhance operational performance. The application of the capabilities of machine learning to this end is of significant worth to businesses that desire to keep up with the increasingly dynamic e-commerce environment. With the growing levels of competition and evolving customer needs, businesses must value the incorporation of intricate analysis into their functions more highly. With the support of the capabilities of machine learning, businesses can gain a deeper understanding of their customers, anticipate market trends, and avert the dangers of fraud.

Literature Review

E-commerce Trends & Consumer Behavior within the USA

According to Agnes et al. (2024), the e-commerce environment within the US has undergone revolutionary change within the last decade due to a myriad of reasons that have together influenced the behavior of the customer base and the way they go about their purchasing decisions. The most notable trend is the growing usage of mobile devices to conduct e-commerce activities. With the omnipresence of smartphones and tablets, shoppers are increasingly browsing, comparing, and purchasing with greater ease of access compared to the past. This trend has also been fueled by the advancements made by mobile payment technologies that have made the payment process smoother and easier to navigate. As per Chandel et al. (2024), the introduction of social media sites has also revolutionized the way shoppers find brands and interact with them. Social commerce that enables the marketing of products directly on social media sites is also on the rise, especially among the younger crowds that are inclined toward visually oriented sites like Instagram and TikTok. The union of e-commerce with social media sites not only increased the customer base of various companies but also equipped them with information regarding the preferences of the customer base.

Another key driver of purchasing behavior online is the growing priority being placed on data-driven marketing campaigns. With the greater accessibility of software to analyze data, companies can gather and review large quantities of customer information to craft their marketing campaigns to target specific groups of individuals. The art of personalized recommendations is a key means by which customer engagement can be stimulated, and sales can be triggered (Chopra et al., 2020). With algorithms that review purchase histories, browsing patterns, and demographic information, companies can deliver personalized product recommendations that resonate with unique individuals. Not only is this personalization boosting the overall quality of the purchasing experience, but it also promotes brand allegiance as shoppers will likely return to sites that know their preferences and deliver them the information they desire (Kalusivalingam et al., 2022)

Furthermore, the COVID-19 crisis accelerated the e-commerce trend even further, with the need to adapt to lockdowns and social distancing forcing consumers to switch to e-commerce. The sudden online purchasing trend saw a greater emphasis on convenience, speed, and trustworthiness, making the need to innovate delivery strategies and customer service imperative for businesses. The trend of subscription models and same-day delivery is a testament to this trend, with companies racing to keep pace with the changing needs of the customer base that is increasingly looking at quick and convenient purchasing experiences. Overall, the confluence of advancements in technologies, the changing needs of the customer base, and the effect of the crisis globally resulted in a dynamic e-commerce landscape within the U.S., with the need to continually research the behavior of the customer base and purchasing patterns to keep the enterprise afloat (Mohaimin et al., 2025).

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Traditional vs. Machine Learning Strategies in Online Purchasing

Kasemrat & kraiwanit (2022), in the e-commerce analysis arena, traditional approaches have historically relied on the use of rule-based systems and static segmentation techniques to provide insights into customer behavior. Although the techniques were a stepping stone to examining the information at hand, they are, by their very nature, bound by their inability to keep up with the nuances of modern customer interactions. Rule-based analysis is often established upon pre-determined criteria that inform the way the information is segmented or interpreted. For instance, the company could segment the customer by simplistic criteria of gender or age without considering the nuanced behavior patterns that guide purchasing behavior. The static approach can generate oversimplified conclusions that are likely to be inaccurate due to the inability to accommodate the dynamic patterns of customer behavior that are driven by a multitude of factors, such as seasonality, trendiness, and personal choice.

Moreover, traditional analysis is not geared to manage the vast amounts of information that e-commerce creates. With growing e-commerce activity, companies are inundated with information that is coming from a myriad of sources like website interactions, social media activity, and sales activity. Traditional means do not have the scalability to manage this information in real-time to deliver quick and well-informed decisions, thus being a limitation. Inefficiency that is associated with hand analysis can translate to lag information that results in companies being reactive rather than proactive with their strategies (Nabi et al., 2024)

In contrast, machine learning represents a paradigm shift regarding the analysis and interpretation of e-commerce data. Machine learning is perhaps best associated with the potential to automate the recognition of patterns and the decision-making itself. With algorithms that learn from data, companies can find insights that are unavailable with standard analytical approaches (Saleem et AL., 2019). For instance, machine learning algorithms can find complex patterns of customer behavior that enable companies to segment their customers dynamically by their purchase behavior, preferences, and interactions. With dynamic segmentation, marketing can be made increasingly personalized since companies can tailor their communications and offers to unique customer profiles.

Additionally, machine learning algorithms can constantly improve with time by learning the newest sets of data inputs and sharpening their predicting skills. In the dynamic e-commerce landscape, where customer preferences change rapidly, this flexibility is very advantageous. With the support of machine learning, businesses can gain a competitive edge by predicting patterns well ahead of time and catering to customer needs at the time of need. In summary, while traditional analysis was a rudimentary approach to analyzing customer behavior, the limitation of rule-driven systems and inflexible segmentation underscores the need to add machine learning methodologies to enhance analytical capabilities within the e-commerce landscape (Tandon et AL., 2024).

Applications of Machine Learning to E-commerce

Ye & Jonilo (2023), reported that the applications of machine learning to the e-commerce sector are broad and complex, giving companies the potential to leverage the value of information to inform predictive analysis and choice-making. An illustrative case is the prediction of sales patterns with the support of historical information. With the analysis of historical sales information, companies can find patterns that inform the prediction of sales to enable them to enhance their stock management decisions and supply chain management. Machine learning algorithms such as time series analysis and regression models can analyze past sales information to find cyclical patterns, seasonality patterns, and external drivers of customer purchasing behavior. Accordingly, companies can predict the change in demand to enable them to plan to meet customer needs while avoiding unnecessary stock and the cost of them (Buiya et AL., 2024).

Another critical e-commerce usage of machine learning is customer segmentation, which can be achieved with clustering and classification methodologies. Segmentation by traditional approaches is normally made based on simplistic criteria, while with the support of machine learning, companies can segment their customers meaningfully based on complex behavior and preferences. For instance, clustering algorithms can segment the customer with the same purchasing behavior so that the company can learn about unique

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

segments that can be addressed with specific marketing campaigns. Besides this, classification methodologies are also able to predict customer behavior, like the purchase of a customer or cart abandonment. With the knowledge of the properties of various customer segments, companies can craft personalized marketing communications and offers that talk to their target audience to boost the rate of conversions and customer satisfaction levels (Zhang et AL., 2023).

As per Sizan et AL. (2025b), Fraud detection is another area where e-commerce has profited heavily with the introduction of machine learning. With growing online payments comes the danger of increased chances of fraud, necessitating robust protection measures for both the company and the customer. Machine learning algorithms can look at the payment details in real time to pick out behavioral patterns that are likely to indicate a case of fraud. With the aid of algorithms like anomaly detection and supervised learning, companies can develop systems to detect fraud that learn with the latest payment information constantly, growing all the more accurately with time. Not only is this a preventative measure to avert fraud, but also a means to win the trust of the customers since they are aware that their payments are being monitored and secured (Islam et AL., 2024).

Research Gaps

Hasan et AL. (2025), found that despite the broad applications of machine learning to the e-commerce sector, several research areas lag that must be addressed. Of key note is the inability of existing models to build real-time adaptive models that can learn to adapt to dynamic consumption patterns. Classical models of machine learning are ideally suited to analyzing past information, but they lag the pace of change that can arise with customer behavior due to external drivers like economic change, seasonality, and sudden changes in customer opinion. With the need to remain adaptive to a dynamic environment that is constantly changing, research into the area of models that can adapt in real-time with continuous learning of the latest information feeds and resultant adjustments to their projections is a priority area that must be addressed. It will enable companies to forecast emerging patterns of consumption to align their marketing strategies and stock management methodologies with the current customer preferences.

Another notable area of research is the need for scalable AI-driven solutions that enhance personalization and anti-fraud capabilities. With e-commerce increasing at a breakneck speed, the volume of information being generated by customer interactions is likely to multiply manifold. Traditional machine learning models could fall short of scaling to accommodate this volume of information, leading to processing delays and insights lagging. Future research will then need to work on scalable algorithms and architecture that can deal with big sets of information with speed and precision. In addition to this, research is also needed to work on finer personalization strategies that go beyond the simplistic segmentation of demographics. With the support of complex machine learning algorithms, companies can work on hyper-personalized interactions that are responsive to the needs of individuals, increasing customer engagement and loyalty (Rana et AL, 2023).

Data Collection and Exploration

Dataset Overview

The synthetic e-commerce dataset comprises a comprehensive collection of simulated transactional data designed to reflect the dynamics of an online retail environment. This dataset includes detailed records of customer transactions, capturing essential information such as transaction IDs, timestamps, product categories, quantities purchased, and total transaction values. Additionally, customer demographics are represented, encompassing attributes such as age, gender, location, and income levels, which facilitate deeper insights into consumer behavior and preferences. The dataset also features product categories that range from electronics to apparel, allowing for diverse analyses of purchasing trends across different market segments. The data sources utilized for this dataset include simulated transactional data generated to mimic real-world scenarios, alongside purchase history and behavioral analytics that track user interactions and engagement patterns, providing a robust foundation for exploring consumer trends and predictive modeling in the e-commerce landscape.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Feature Selection

S/No.	Feature	Description
01.	Transaction ID	A unique ID per each transaction enables tracking of a purchase by a certain customer within the database.
02.	Customer ID:	Unique ID per customer to aid the build-up of purchase behavior and usage patterns by user
03.	Transaction Date	Date of the purchase and time of the purchase: it is convenient to look at patterns of purchase by time interval, seasonality, and peak purchase seasons.
04.	Product Category	Categorization of the purchased products, e.g., electronics, clothes, or housewares, to permit analysis of the category performance of the sales
05.	Quantity Purchased	Units purchased per purchase, providing information regarding customer purchasing behavior and the popularity of specific products
06.	Total Transaction Value	The aggregate value of spending per transaction can be applied to the analysis of customer spending behavior and the generation of revenues.
07.	Customer Age	Customer age to present a profile of the customer by age group to segment marketing efforts appropriately
08.	Customer Gender	Customer gender can be applied to discover information regarding gender groups of shoppers with distinct purchasing behavior patterns.
09.	Location	The customer's location of residence to gain regional trend analysis and preferences, with specific marketing campaigns to target them accordingly
10.	Previous Purchase Record	a summary of the customer's past purchase behavior that can be leveraged to deliver personalized offers and reward customer loyalty

Data Preprocessing

The Python code performed preprocessing of the data of a classification problem that is likely to be sales-related or advertising-related. It begins with importing the necessary libraries like pandas, matplotlib, seaborn, and scikit-learn. It then loads the data into a CSV (assumed to be the case), converts the column 'Transaction Date' into objects of the datatype date-time, and generates new features like 'Year,' 'Month,' 'Day,' and 'Day of Week.' Categorical features like 'Customer ID,' 'Product ID,' 'Category,' and 'Region' are also encoded with the support of Label Encoding. It treats the missing values by replacing them with 0. It then selects a feature set and generates a target variable, 'High Revenue,' according to if the value of the column 'Revenue' is greater than the column's median value. It then splits the data into a training set and a testing set and standardizes the numeric features with the support of Standard Scaler before the modeling (which is not present in the image).

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial initial step of the research process that involves the usage of both visual and statistical techniques to look at and summarize the key characteristics of a dataset. Inspecting the information at the various levels serves to unveil patterns, discover outlying observations, verify assumptions, and generate hypotheses to lead the analysis to proceed. It helps the investigator to gain a deeper sense of the overall organization of the information, potential relationships between the involved variables, and potential quality issues with the information that needs to be addressed before continuing to the next stages of more rigorous modeling or testing of hypotheses.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Daily Revenue Trend

The Python program used the Pandas and Matplotlib-Seaborn libraries to examine and graph daily revenue patterns within a transactional database. It first converts the "Transaction Date" column to date objects to allow proper time-series analysis to occur. Following the establishment of a visually pleasing theme with Seaborn, the program creates a daily revenue line graph. It groups the data by the date component of the "Transaction Date" column and aggregates the "Revenue" by the day. It then graphs this aggregate information with a marker at each data point and a blue line between them. The graph includes a descriptive title, axis titles, a rotated date of the x-axis to improve readability, a grid to enhance visualization, and a tight layout to avoid overlapping elements. It then displays the graph with the call to plt.show(). Essentially, this program successfully graphs the daily revenue trend by time to allow analysts to notice patterns and possible abnormalities.

Output:

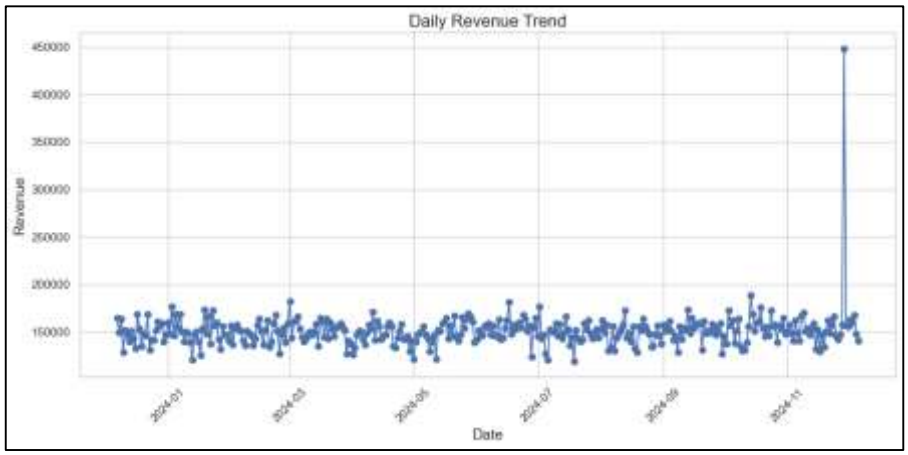

Figure 1. Daily Revenue Trend

The histogram of the "Daily Revenue Trend" within the time frame displays a predominantly stable trend of revenues with daily fluctuations between the 15,000 to 20,000 range for the majority of the days, representing stable sales activity. An outlier is strongly noticeable on the date labeled 2024-11, with the sudden jump of the revenue to a peak of about 450,000, implying a major event or a special campaign that resulted in a record number of sales on that specific day. The sudden spike is a striking contrast to the otherwise stable trend of the revenues, illustrating the potential of the effect of specific marketing campaigns, seasonality of sales, or the introduction of a new product that can generate a significant jump in revenues. The overall trend reflects the need to observe the daily sales variability to find the key drivers of the revenues to align the marketing strategies accordingly in the future.

Top 10 Products by Revenue

The code script in the Python program employed the usage of Pandas to calculate the overall revenue of each product with group by () and sum() functions and rank the revenues by highest to the best performers with sort values() and head(10). The derived information saved to the variable product revenue is then plotted with a bar graph with kind='bar'. The graph aesthetics are included with a title, the title of the axes, and the rotation of the x-axis to enhance readability with rotation (x). The color of the bars is teal, while the call to plt.tight_layout() ensures that all elements are within the figure limits. The final output of the produced bar graph is displayed with the call to plt.show().

Output:

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

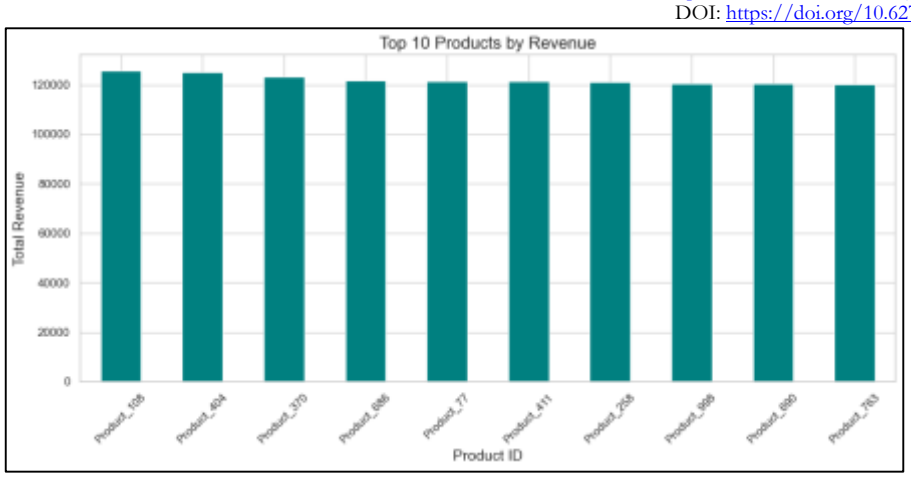

Figure 2. Top 10 Products by Revenue

The histogram of the "Top 10 by Revenue" reveals that the selected products—their corresponding Product ID—report very identical total amounts of revenue, with all of them being very much the 120,000 value. The similarity of the results reflects that the products are among the best-performing products of the e-commerce database with strong customer base and well-performing marketing campaigns. Of note are the products such as the Product_109, the Product_404, and the Product_370 that are the key drivers of overall sales with a significant value to the overall revenue. The similarity of the amounts of the revenue among the best-performing 10 products reflects a well-balanced performance with the customer base, revealing a broad range of preferences that translate into stable sales across a number of the best-performing products rather than a best-seller. The information can inform stock management and marketing strategies to sustain stock levels and marketing campaigns of highly revenue-generating products.

Revenue by Region

The Python script used Pandas and Matplotlib to visualize revenue distribution across different regions. It calculates the total revenue for each region by grouping the data by 'Region' and summing the 'Revenue' values. The resulting region revenue data is then sorted by revenue in ascending order. A horizontal bar chart (kind='barh') is generated to display this data, with each bar representing a region and its length corresponding to the total revenue. The bars are colored purple for visual distinction. The chart is enhanced with a title, axis labels, and a tight layout to prevent overlapping elements, ensuring a clear and informative visualization of region-wise revenue distribution.

Output:

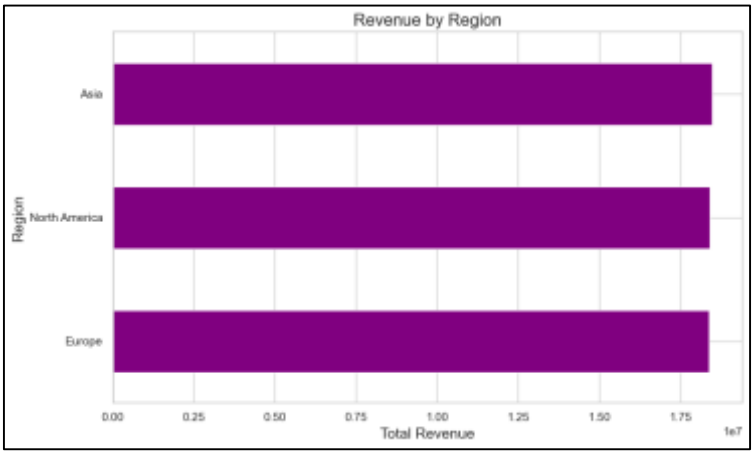

Figure 3. Displays Revenue by Region

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

The histogram illustrating "Revenue by Region" reveals distinct revenue contributions from three major geographic areas: Asia, Europe, and North America. Notably, Europe leads with a total revenue of approximately 1.75 million, indicating its status as a key market for the e-commerce business. Asia follows closely, also generating significant revenue, while North America lags, contributing around 1 million in total revenue. This disparity suggests that European and Asian markets may exhibit stronger consumer engagement or a broader product appeal compared to North America. The data underscores the importance of regional marketing strategies, as businesses can tailor their approaches to capitalize on the higher revenue potential in Europe and Asia while exploring opportunities to boost sales in the North American market. Additionally, understanding the factors driving revenue in these regions can inform decisions regarding product offerings and promotional efforts to enhance overall sales performance.

Showcases Correlation Heatmap of Key Features

The Python script generated a correlation heatmap to visually represent the relationships between several key features within a dataset. It begins by selecting specific columns relevant to the analysis, including 'Units Sold', 'Discount Applied,' 'Revenue,' 'Clicks,' 'Impressions,' 'Conversion Rate,' and 'Ad Spend,' and calculates the pairwise correlation matrix using the .corr() method. This matrix is then visualized as a heatmap using Seaborn's heatmap() function. The heatmap displays the correlation coefficients with annotations (annot=True) formatted to two decimal places (fmt=".2f") and employs the cool, warm color scheme to intuitively represent positive and negative correlations. The square=True argument ensures the heatmap is displayed as a square, and a title is added for clarity. Finally, plt.tight_layout() adjusts subplot parameters for a tight layout, and plt.show() displays the generated heatmap, providing a clear visual representation of the relationships between the chosen variables.

Output

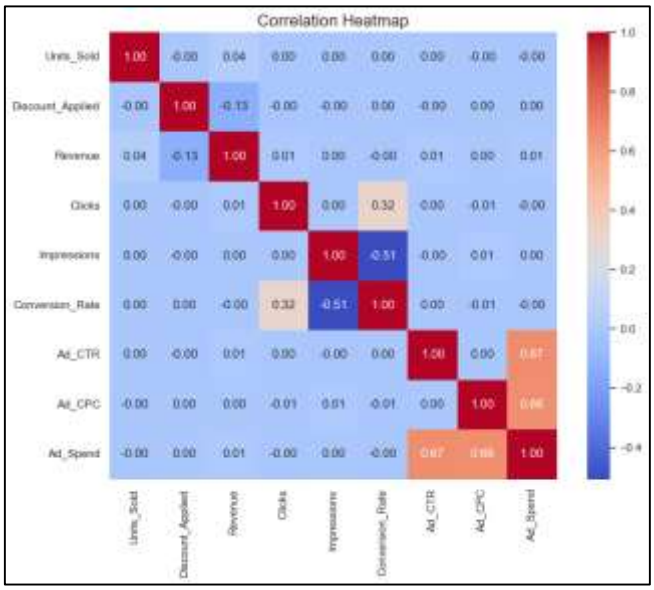

Figure 4. Showcases Correlation Heatmap of Key Features

The correlation heatmap provides a comprehensive overview of the relationships between the metrics of e-commerce performance. "Units Sold" is the most highly correlated with "Revenue," both with a 1.00 value of correlation, meaning that with units sold comes the corresponding increase in revenue. "Ad Spend" is also highly positively correlated with "Revenue" at 0.67, meaning that increased spending on advertising is associated with increased sales, validating the effectiveness of the advertising campaigns. "Clicks" is also revealed to have a negative correlation of -0.51 with "Conversion Rate," meaning that increased clicks may generate traffic but do not necessarily translate into purchase activity, showing potential with the sales funnel or the desirability of the products. The less strong correlations of "Impressions" with other metrics

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

indicate that the increased presence of ads is no guarantee of sales, showing the need to target marketing efforts accordingly. Overall, this heatmap can provide key performance metrics and relationships that can inform data-driven decisions to improve marketing campaigns and the outcome of sales.

Sales Distribution by Category

The Python code snippet plotted the sales distribution by various product categories. It uses Pandas to group the data by 'Category' and find the total 'Units Sold' per category with the aid of .groupby() and .sum(). The category sales data is then sorted in descending order to easily pick out the best-performing categories. It creates a bar graph with the aid of .plot(kind='bar'), with each category represented by a bar and the bar height equal to the number of units sold. The bars are filled with the color coral to add aesthetic value to the graph. It is beautified with a title, the title of the axes, and the rotation of the x-axis to improve readability. It uses plt.tight_layout() to have a clean layout and plt.show() to produce the final graph that well represents the sales distribution by various product classes.

Output:

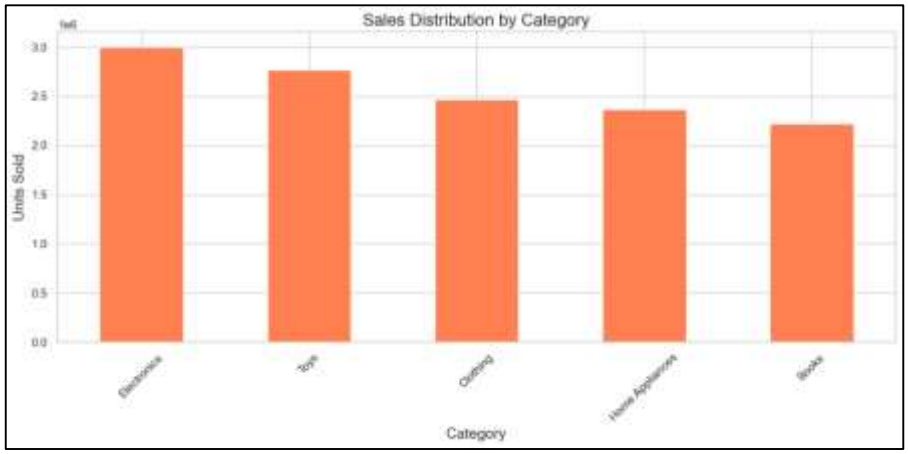

Figure 5. Sales Distribution by Category

The histogram of "Sales by Category" showcases the number of units sold within five separate categories: Electronics, Toys, Clothing, Home Appliances, and Books. Electronics are the best-performing category, with about 3 million units sold, representing strong customer appetite and possibly innovative products. Toys are a very close second with strong market performance, driven by family-oriented buys. Both Clothing and Home Appliances have the same number of sales, slightly less than 2 million units sold, representing stable customer appetite for the necessary consumer products. In contrast, the Books category is also significant but less so, with about 1.5 million units sold being the closest to this category. The spread reflects the need to prioritize marketing efforts to the strong-performing category of Electronics and Toys while also considering ways to revitalize the Books category to generate sales. Overall, the information is used to inform stock management decisions and specific marketing campaigns to generate the highest possible sales within all categories.

Ad Performance: CTR vs. CPS

The Python code generated a scatter plot to compare the Click-Through Rate (CTR) with the Cost Per Click (CPC) of ads with other dimensions of Region and Ad spending. It uses the scatterplot() function of Seaborn to plot 'Ad CTR' vs. 'Ad CPC,' with the area of the marker varying concerning 'Ad Spend' and the color of the marker varying concerning the 'Region.' Alpha is the added translucency to the marker, while the range of the marker areas is established by the size variable. It is supplied with a title, the name of the axes, and a legend that is placed outside the figure area to avoid confusion. The proper spacing is ensured by the call to plt.tight_layout(), while the scatter plot is displayed by the call to plt.show().

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Output:

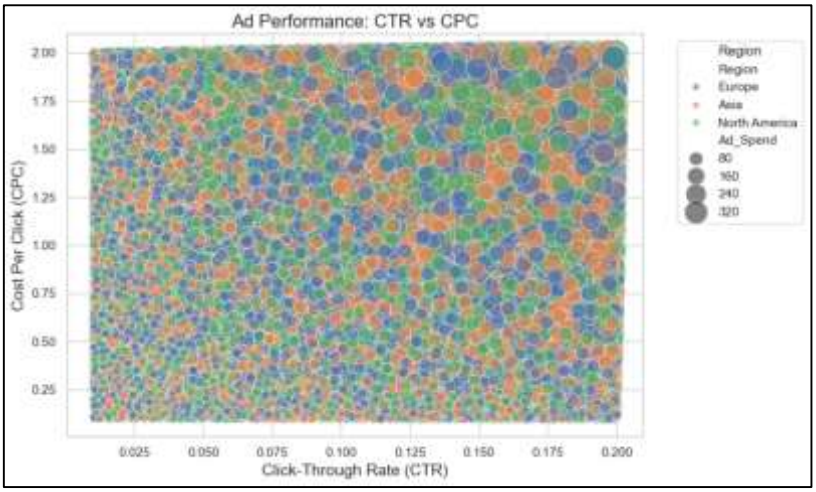

Figure 6: Ad Performance: CTR vs. CPS

The scatter plot titled "Ad Performance: CTR vs CPC" provides a detailed visualization of the relationship between Click-Through Rate (CTR) and Cost Per Click (CPC) across different regions—Europe, Asia, and North America. Each point represents an advertisement, with the size of the circles indicating the level of ad spend, allowing for a nuanced analysis of ad performance. Notably, the data reveals that advertisements from Europe generally exhibit higher CTR values compared to those from Asia and North America, suggesting more effective engagement in that region. Conversely, CPC values remain relatively similar across all regions, indicating that while European ads may perform better in terms of clicks, the cost per click does not significantly differ. The distribution of points shows a cluster of ads with low CTR and varying CPC, highlighting potential inefficiencies in certain campaigns that may require optimization. Overall, this visualization aids in understanding regional ad performance dynamics and informs future advertising strategies to enhance effectiveness and return on investment.

a) Distribution of Click-Through Rate (CTR)

The Python script looked at click-through rates (CTR) by calculating their distribution and graphing them. It first calculates CTR as a percentage by dividing the column of 'Clicks' by the column of 'Impressions' and assigning the value to a new column titled 'Click Through Rate.' It then generates a 30-binned histogram with the histplot() function of the Seaborn module to graph the value of the CTR distribution. The argument of kde=True is added to include a kernel density estimate to enable a smooth graph of the distribution to be plotted. The green color is filled into the 30-binned histogram. It is then finished with a title, titles of the axes, and a tight layout to avoid overlapping elements before being displayed with the call to plt.show(), providing a sense of the normal range and frequency of the click-through rates.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Output:

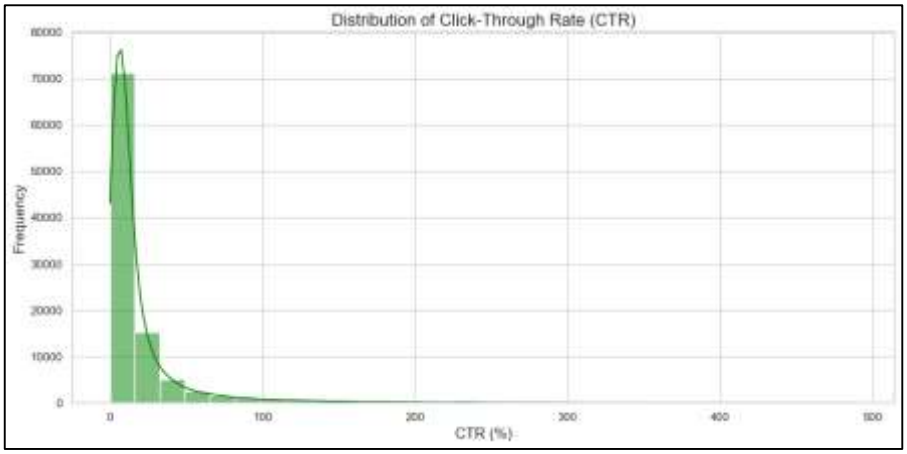

Figure 7. Distribution of Click-Through Rate (CTR)

The histogram of the "Distribution of Click-Through Rate (CTR)" is skewed heavily with the larger number of CTR values at the lower end of the scale, illustrating that the vast majority of ads have minimal engagement. It is at very low CTR percentages that the frequency is highest, illustrating that the campaigns struggle to generate a lot of clicks per impression. Not many ads have significantly higher CTR values with a very drawn-out tail of the distribution that represents exceptional performance by a small number of campaigns. The evidence is illustrated by the struggle that advertisers have to generate engaging ads, with the vast number of ads failing to generate meaningful click rates. The graph reinforces the need to improve the ad copy and targeting approaches to improve CTR while also considering that a small number of exceptional campaigns can have a skewed impact on overall performance metrics. Overall, the evidence is that a more deliberate approach is needed to improve engagement overall.

Monthly Revenue Trend

The Python script visualizes the trend of monthly revenue over time. It begins by extracting the month from the 'Transaction Date' column and storing it in a new 'Month' column using .dt.to_period('M'). It then calculates the total revenue for each month by grouping the data by 'Month' and summing the 'Revenue' using .groupby() and .sum(). The resulting monthly_revenue data is plotted as a line chart with markers using .plot(kind='line', marker='o'), with a dark orange color. The chart includes a title, axis labels, rotated x-axis labels for readability, and a grid for better visualization. Plt.tight_layout() ensures a clean layout, and plt.show() displays the final plot, effectively showcasing the monthly revenue trend.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Output:

Figure 8. Monthly Revenue Trend

The histogram of the "Monthly Revenue Trend" displays the performance of the revenues year-round with a predominantly stable trend with minimal variability of the month-wise revenues. It is worth noticing that the revenues start at around 4.5 million at the end of the year and are at a similar value at the beginning of the year with minimal increases and drops, showing regular sales activity. However, a significant decline is observed at the end of the year at around 1 million, which can be due to seasonality, stock issues, or market conditions. The sharp decline at the end of the year can necessitate a reassessment of the sales strategies and stock management to align with the needs of the customer at the time of peak seasons of purchase. Overall, the trend displays the need to monitor constantly with adjustments to both seasonality patterns and market conditions to maximize the revenues year-round.

Revenue Contribution b Product Category

The Python script generated a pie chart to visualize the contribution of each product category to the total revenue. It used Pandas to group the data by 'Category' and calculate the sum of 'Revenue' for each category. The resulting category revenue data was then plotted as a pie chart using .plot(kind='pie'). The autopsy parameter formats the percentage labels on each slice to one decimal place. The start angle rotates the first slice by 140 degrees for better presentation, and cmap='viridis' applies the viridis colormap for distinct slice colors. The chart includes a title and omits the y-axis label since it represents revenue contribution. Plt.tight_layout() ensures a clean layout, and plt.show() displays the final pie chart, effectively illustrating the proportion of revenue generated by each product category.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

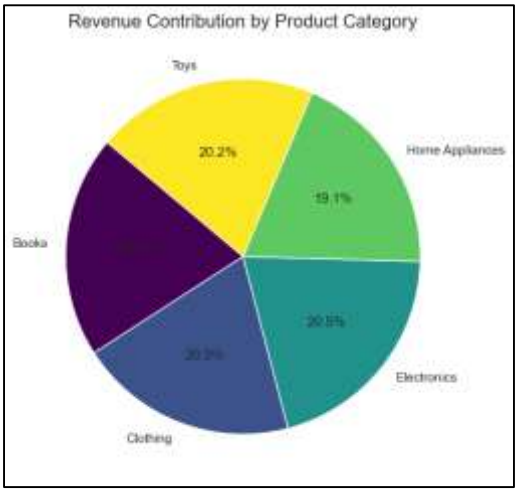

Figure 9. Revenue Contribution b Product Category

The pie chart "Revenue Contribution by Product Group" represents the comparative revenue contributed by the five distinct groups of Toys, Home Appliances, Electronics, Clothing, and Books. All the groups have a fairly equal contribution, with Electronics contributing the highest at 20.5%, followed by Toys and Clothing at 20.2%. Home Appliances have a slightly lower contribution of 19.1%, while Books lag at 20.0%. The equal split is a sign of a diversified stream of revenue that lessens the effect of a single category upon which the company is highly dependent, cushioning the effect of turbulence in the market. The similarity of the contributions is a sign that all the groups have a contributing role to the overall revenue, necessitating a well-balanced portfolio of products. To obtain the highest possible potential of sales, companies must have specific marketing strategies that leverage the strengths of each group while also examining the potential to enhance the revenue of the weaker groups.

Impressions vs. Conversion Rate

The Python program generated a scatter plot with a regression line to plot the relation between 'Impressions' and 'Conversion Rate.' It employs the replot () function of the Seaborn module to have a linear regression model of the data automatically plotted with the scatter plot of the two variables with a regression line superimposed upon them. scatter_kws is employed to modify the scatter points' transparency (alpha) to 0.5, while the color of the regression line is made red by the use of line_kws. The title of the graph and the titles of the axes are included to add readability to the graph, with the title of the y-axis declaring that the 'Conversion Rate' is a percentage. Plt.tight_layout() is invoked to add proper spacing between the components of the graph, while the graph is displayed by the call to plt.show().

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Output:

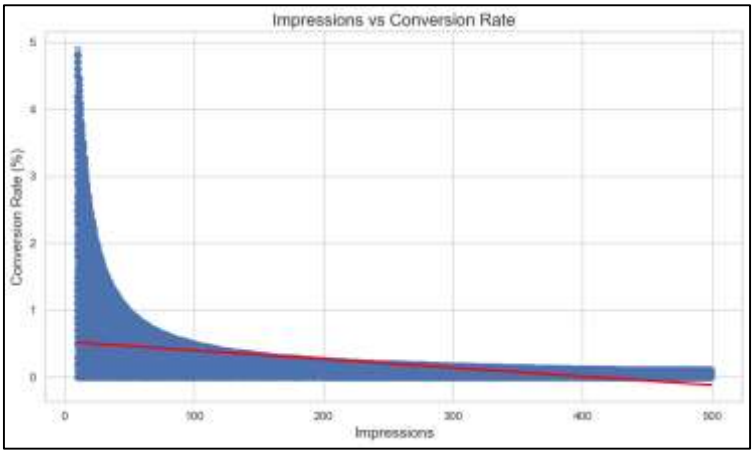

Figure 10. Impressions vs. Conversion Rate

The scatter graph "Impression vs Conversion Rate" reflects the interaction between impressions and the resultant rate of conversions with a strong negative trend. With a rise in impressions, the rate of conversions significantly drops, with the highest rates of conversions being mostly at the lower end of impressions. It is a sign that while a lot of impressions can have the first glimpse at the time of a click, they do not necessarily lead to corresponding conversions with signs of inefficiency of the sales funnel. It is supported by the trend of the red trend line that reflects that beyond a certain number of impressions, the rate of conversions levels off to a value of zero with diminishing returns to impressions. It is a sign that advertisers have to improve their targeting and engagement strategies to not simply increase the number of impressions but the quality of interactions to improve the rates of conversions positively.

Distribution of Discounts Applied

The Python fragment plotted the distribution of applied discounts. It employed the heist plot () function of the Seaborn module to produce a histogram of the 'Discount Applied' column of a Pandas Data Frame called data. The histogram is supplemented with a Kernel Density Estimate (set to kde=True) to present the smoothened distribution curve. The color of the histogram bars is salmon, with the number of bins being 30 to provide a detailed representation. The figure is 12 by 6 inches in size. The title of the visualization is that it represents the applied discount distribution with the proper title, the x-axis title (representing the discount percentage), and the y-axis title (representing the frequency). In the end, the subplot parameters are made to have a tight layout by plt.tight_layout(), and the produced histogram is shown by plt.show() to provide information about the frequency of the applied discounts and their range.

Output;

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

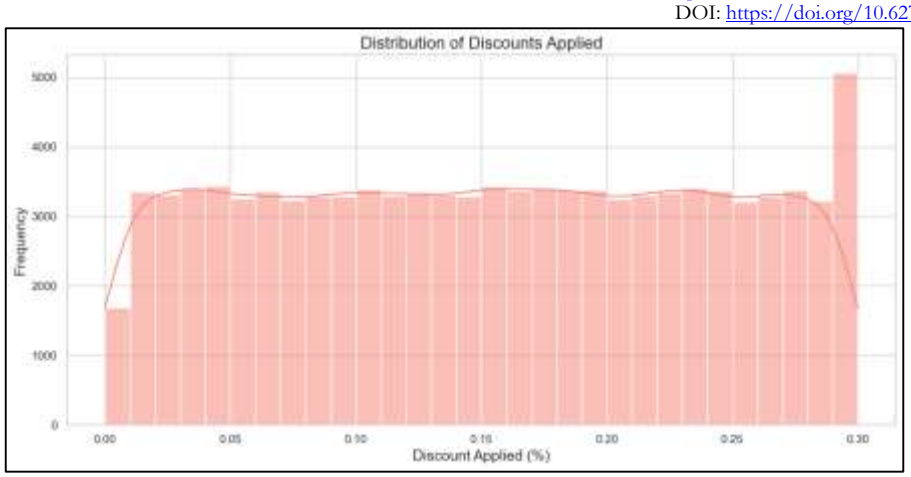

Figure 11. Portrays the Distribution of Discounts Applied

The histogram "Distribution of Discounts Applied" represents the frequency of the range of the discount rates applied to the transactions, with a concentration of the discounts predominantly at the lower end of the range between 0% and 10%. Notable is a sharp peak at the 30% discount with a little less than 5,000 occurrences, which represents the selective usage of the big discounts to push sales or offload stock. The overall shape of the distribution is a gradual fall off of frequency with increasing levels of the discount that represents that while the small discounts are being frequently applied, the big discounts are being applied sparingly. The information represents the effectiveness of the intermediate discounts to push purchases while also representing the potential of deeper discounts to generate customer excitement at the time of specific campaigns. Companies can consider analyzing the impact of discount strategies on overall sales performance and customer retention to improve the pricing models going forward.

Ad Spend vs. Revenue by Region

The Python script creates a scatter plot to explore the relationship between 'Ad Spend' and 'Revenue' while also considering the influence of 'Region' and 'Ad CTR.' Using Seaborn's scatterplot() function, it plots 'Ad Spend' on the x-axis and 'Revenue' on the y-axis. The hue parameter colors the points based on the 'Region,' allowing for the visual distinction between regions. The size parameter scales the points based on 'Ad CTR,' making points with higher 'Ad CTR' appear larger. The sizes argument sets the range for the point sizes, and alpha controls the transparency of the points. A title and axis labels are added for clarity, and a legend is placed outside the plot area using bbox_to_anchor for better readability. Plt.tight_layout() ensures that the plot elements fit within the figure boundaries, and plt.show() displays the generated scatter plot, providing insights into how 'Ad_Spend,' 'Revenue,' 'Region,' and 'Ad_CTR' interact. Note that the palette argument within SNS.scatterplot is incomplete in the provided code snippet and would require a color palette to be fully functional.

Output:

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

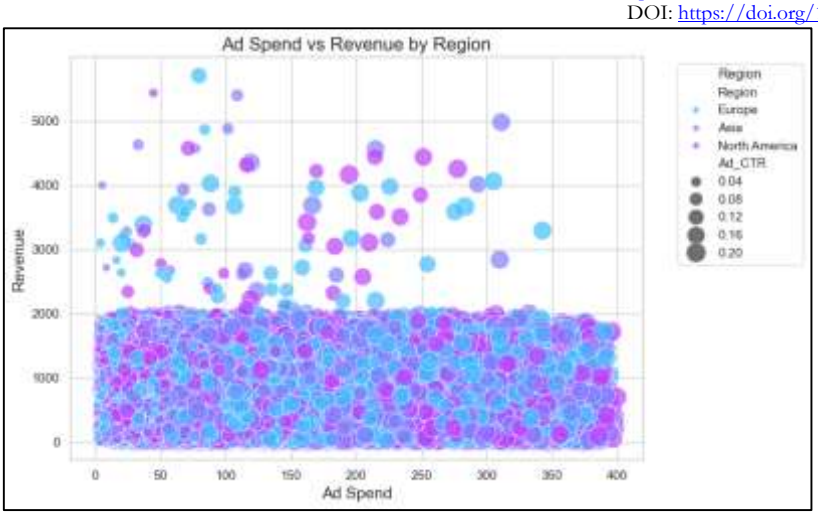

Figure 12: Ad Spend vs Revenue by Region

The scatter graph "Ad Spend vs Revenue by Region" displays the correspondence between advertising spending and the resultant revenue in three separate regions of Europe, Asia, and North America. Every data point is a separate advertisement, with the circles representing the Click-Through Rate (CTR), giving a measure of the effectiveness of each advertisement. From the graph, the trend is that increased spending on ads is normally associated with increased revenue, while the spread of the data points reflects a lot of variability of the resultant revenues by region. Of note is that the range of revenues is greater at the same spending levels of ads by the region of Europe compared to the case of Asia and North America, implying varying market conditions or responses by the consumers. The varying circles also indicate the significance of CTR since the bigger circles correspond to greater engagement, which is key to maximizing the resultant revenue. Overall, the graph reflects the imperative of targeted advertising campaigns specific to regional behavior and the need to maximize spending and engagement to improve overall advertising performance.

Methodology

Feature Engineering

In the feature engineering phase, we identified and selected key predictive variables essential for enhancing model performance. According to Ray (2025), the primary features include purchase frequency, which quantifies how often customers make purchases, thus providing insights into their engagement levels; spending behavior, which encompasses average transaction values and spending patterns to understand customer financial commitment; and customer lifetime value (CLV), an estimate of the total revenue a customer generates throughout their relationship with the business. As per Islam (2025), these features are critical as they directly influence customer retention and profitability. To further enhance our database, we derived new features by considering time series and behavioral patterns. For example, we computed the recency of a customer's last purchase date to indicate the time lapse between the customer's last purchase date and the current date and seasonality features to observe patterns within certain time intervals like festivals or sales seasons. We also derived features like customer segmentation by purchase behavior and churn propensity to forecast the chances of a customer to stop purchasing. These derived features allow us to have finer-grained insights so that models can learn intricate relationships and patterns of customer behavior over time.

Model Training and Selection

For model choice, we applied various machine learning algorithms specific to our needs of predicting sales, customer segmentation, demand forecast, and fraud detection. Random Forest, Logistic Regression, and K-Neighbors Classifier are the models selected by the analyst. Random Forest was selected for its robustness and ability to handle non-linear relationships without extensive parameter tuning, making it ideal

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

for complex datasets with numerous features. Rahman et al. (2023), argued that Logistic Regression, despite its simplicity, provides interpretability and efficiency for binary classification tasks such as fraud detection. It allows for clear insights into the influence of various features on the probability of fraud. Lastly, the K-Neighbors Classifier was chosen for its effectiveness in clustering similar data points, which is particularly useful for customer segmentation, as it can identify groups of customers with similar purchasing behaviors. The selection of these models was guided by the characteristics of the dataset, including feature types and distribution, ensuring that we utilized algorithms best suited for our predictive tasks.

Model Optimization and Performance Analysis

To enhance model accuracy, we conducted hyperparameter tuning, employing techniques such as Grid Search and Random Search to identify optimal parameters for each model. This process is crucial as it fine-tunes the models to better fit the training data without overfitting. We also cross-validated to check that the models would work well with unknown data by dividing the dataset into training sets and a distinct set of sets of validations several times. It helps the testing of the performance of the model together with its robustness. We also performed a feature importance analysis to estimate the effect of each feature on the model's predictions. Having the most important features helps us to enhance the models again and to make sensible judgments about the areas of possible improvement within the feature set (Ray et al., 2025).

Evaluation Metrics

For evaluating the performance of the models, we used a suite of metrics specific to the task at hand. For the case of fraud detection, the metrics included were accuracy, precision, recall, F1-score, and ROC-AUC. According to Islam et al. (2025), these metrics are a complete representation of the performance of the model to correctly classify fraudulent transactions with minimal false negatives and false positives. In contrast, to predict sales and demand, we have used Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to estimate continuous prediction quality. RMSE is very sensitive to the presence of outliers, while MAE provides a straightforward representation of the average error of the predictions with a well-balanced performance view. In summary, the performance metrics combined enable us to measure the effectiveness of the models appropriately and guide data-driven decisions to enhance the quality of the predictions.

Results and Analysis

Model Performance Comparison

K-Nearest Neighbors (KNN) Classifier Modelling

This Python script program employed the sci-kit-learn module to develop a K-Nearest Neighbors (KNN) classifier. It defines a function knn classifier that takes training and testing data (features and target variable) as the argument. In the function, a K-Neighbors-Classifier is established with the argument n-neighbors=5 to include the five closest neighbors to make a prediction. It trains the classifier with the training data (X-train, y-train) with the invocation of the fit() method. It predicts the output on the test set (X-test) with the invocation of the predict() method, with the results being stored in the variable y-pred. It then prints a classification report with the precision, recall, F1-score, support of each class, and overall accuracy with the invocation of the accuracy score() function. The KNN-classifier function is then called with the training and testing data. This program successfully implements the training of a KNN classifier on a certain dataset with the corresponding testing of the classifier.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Output:

Table 1.KNN Classification Report

```
--- K-Nearest Neighbors (KNN) Classifier ---
              precision    recall  f1-score   support

           0       0.51      0.50      0.51      9998
           1       0.51      0.51      0.51     10002

    accuracy                           0.51     20000
   macro avg       0.51      0.51      0.51     20000
weighted avg       0.51      0.51      0.51     20000

Accuracy: 0.51
```

The table reflects the K-Nearest Neighbors (KNN) Classifier performance metrics of its recall, precision, F1-score, and support on two classes with the 0th and 1st designation, respectively. For both classes, the recall is 0.51 while the precision is 0.50, respectively, providing the classifier with a 0.51 F1-score that is a well-balanced although comparatively low performance at instance classification. The measures of support are 9,998 of the 0th class and 10,002 of the 1st class, respectively, showing a fairly equal split between the two classes. Overall, the classifier is 0.51 accurate, a performance that is slightly higher than random guesswork, showing possible weaknesses within the model to discriminate between the classes well. The performance is a sign that the classifier could need to undergo a change of approach or be optimized to improve the outcome of the classification.

Random Forest Classifier Modelling

The Python code trained and tested a Random Forest Classifier with the scikit-learn module. It creates a function random forest_ classifier that accepts training and testing sets (features and labels). Within the function, a Random Forest Classifier with 100 estimators (n-estimators=100) is defined with a set random state to allow reproducibility (random-state=42). The training data is then applied to the model with the fit() function to train the model. Predictions are made with the test data with the predict() function, and the performance of the classification is assessed with a classification report and accuracy score that are printed to the console. The target variable (y-train, y-test) is properly shaped with the values.ravel() call. The random forest classifier function is then called with the training and testing sets to train the Random Forest model, test the performance of the Random Forest model, and print the performance to the console. SVC and Logistic Regression are imported classifiers that are not called within this code segment.

Output:

Table 2. Random Forest Classification Report

```
--- Random Forest Classifier ---
              precision    recall  f1-score   support

           0       0.56      0.60      0.58      9998
           1       0.57      0.54      0.55     10002

    accuracy                           0.57     20000
   macro avg       0.57      0.57      0.57     20000
weighted avg       0.57      0.57      0.57     20000

Accuracy: 0.57
```

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

The table summarizes the Random Forest Classifier performance measures with its precision, recall, F1-score, and support of two classes, 0 and 1. For the 0th class, the precision is 0.56 while the recall is 0.60, with the resultant being the 0.58 F1-score that reflects a moderate ability to correctly classify the occurrences of this class. For the 1st class, the trend is the same, with the same precision and recall rates to deliver the same 0.58 F1-score. The support measures have 9,998 occurrences of the 0th class and 10,002 of the 1st class, with a well-balanced split being reflected. Overall, the classifier is 0.57 accurate, with a performance that is better compared to the KNN model. From this performance, the Random Forest Classifier is revealed to have a certain potential to discriminate between the two classes with a need to improve, especially the precision and the F1-score, to deliver a strong classification outcome.

Logistic Regression Modelling

The code script in the Python program defined and used a function named logistic regression classifier to train and test a Logistic Regression model. The function takes training and testing sets (features and target variables) as parameters. In the function, a Logistic Regression model is defined with a certain random state to reproduce the results and a max iteration of 1000. It trains the model with the training set (X-train, y-train) by calling the fit() function. It predicts the results on the testing set (X-test) by calling the predict() function, and the performance of the classification is evaluated by the print of the classification report (precision, recall, F1-score, support) and the overall accuracy by the call of the accuracy score() function with two digits of the format %.

Output:

Table 2. Logistic Regression Results

```
--- Logistic Regression Classifier ---
              precision    recall  f1-score   support

           0       0.53      0.53      0.53      9998
           1       0.53      0.53      0.53     10002

    accuracy                           0.53     20000
   macro avg       0.53      0.53      0.53     20000
weighted avg       0.53      0.53      0.53     20000

Accuracy: 0.53
```

The table above is a summary of the Logistic Regression Classifier performance measures with the precision, recall, F1-score, and support of two classes being represented by 0 and 1. Both classes have the same value of all the measures with the precision, recall, and the F1-score, all at 0.53, with a regular but low ability to correctly classify the instances. The support measures indicate 9,998 of the instances of class 0 and 10,002 of class 1, with an equal split of the data between the two classes. Overall, the classifier is 0.53 accurate, with a little higher than a random guess, with the effectiveness of the model to split the two classes being very low. The performance reflects the need to improve the model or to look at other approaches to the modeling to enhance the ability to forecast.

Comparison of All Models

The Python code implemented a comparison of the performance of the Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression classification models. It defined independent functions to train and test each of the models, with the results (accuracy and confusion matrix) being stored in a dictionary called model results. Each of the functions trains the corresponding model, predicts the test set, calculates the accuracy with the assistance of an accuracy score, generates a confusion matrix with a confusion matrix, prints a classification report, and then stores the accuracy and confusion matrix to the model_results dictionary. Once all the models have been trained and tested, the program prints the results graphically. It

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

first extracts the accuracies of each of the models to a bar graph to compare them graphically. It then iterates the model_results dictionary to generate and print a confusion matrix heatmap of each of the models to provide a graphic representation of their performance concerning the actual and predicted labels. This is a comprehensive way of comparing the strengths and weaknesses of the various classification models.

Outpu:

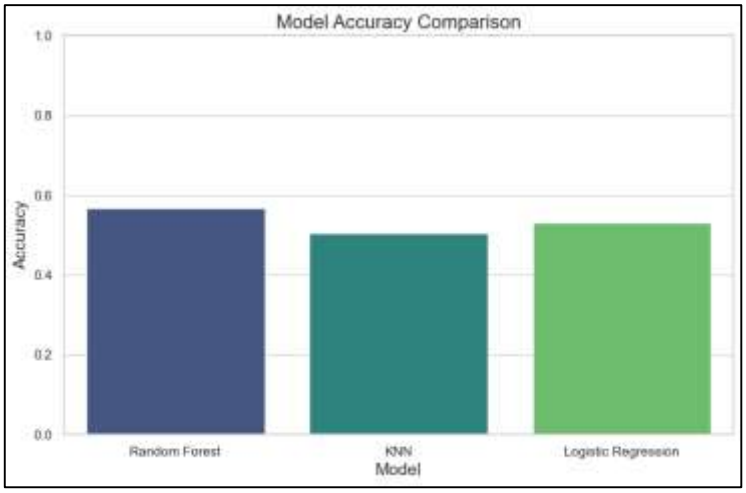

Figure 13. Model Accuracy Comparison

The histogram "Model Accuracy Comparison" pictorially displays the comparison of the accuracies of the classifiers Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. The heights of the bars indicate the accuracies of the models, with Random Forest being the highest at approximately 0.57, followed by KNN at approximately 0.51, while Logistic Regression is at 0.53, being the least among the two models. The varying heights of the bars pictorially display the comparative performance of the models, with Random Forest being well ahead of the other two models. The pictorial representation of the variability of the performance of the classifiers is a reminder of the caution to exercise while selecting the models concerning the measure of accuracy to improve the outcome of the analysis to a greater degree. Overall, all the models have comparatively low accuracies, with the histogram revealing that Random Forest is the best among the three models for this specific dataset.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

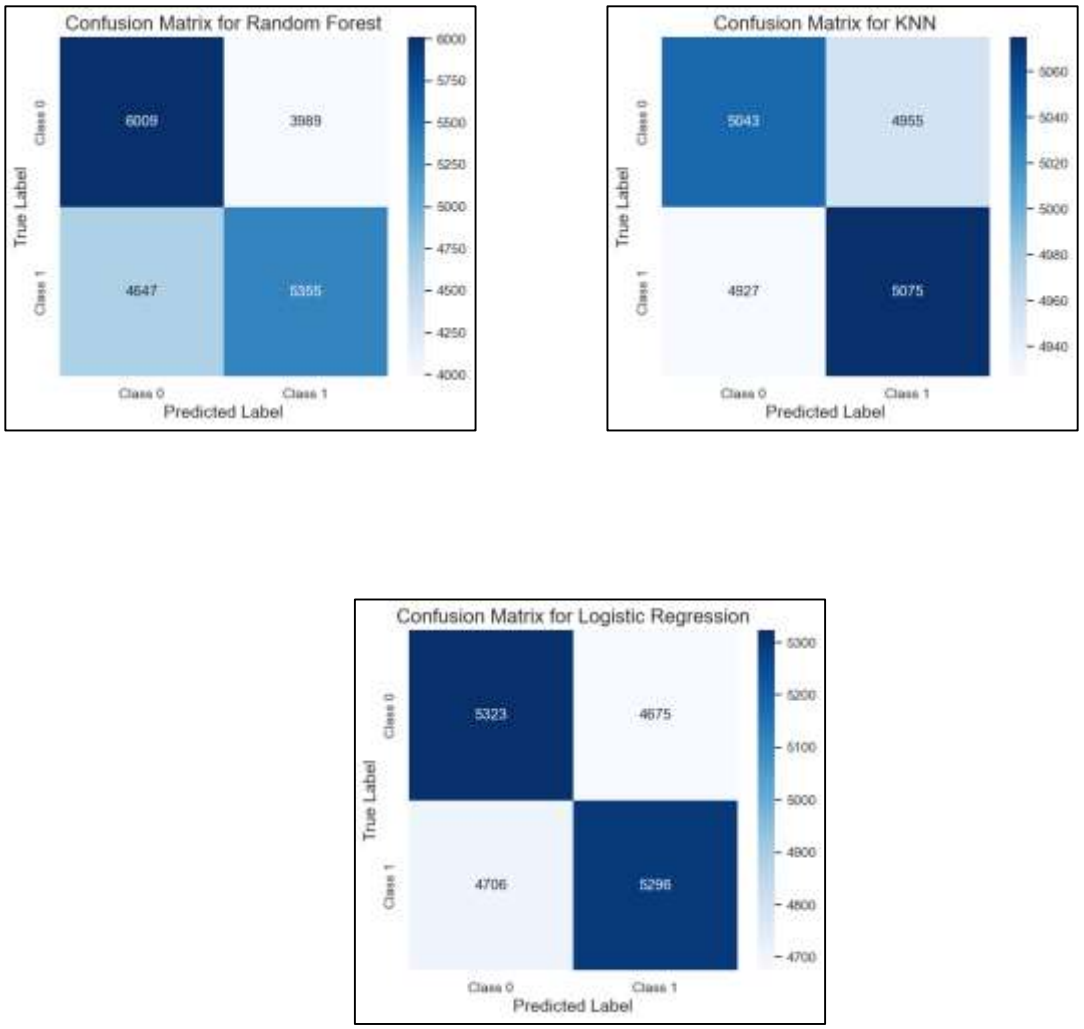

Figure 14. Confusion Matrix Comparison

The confusion matrices for the Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression classifiers provide a detailed view of their classification performance across two classes, labeled 0 and 1. The Random Forest model correctly predicted 6,009 instances of class 0 but misclassified 3,589 as class 1, while 4,647 of class 1 were correctly identified, and 5,535 were misclassified as class 0. The KNN classifier shows a similar trend, with 5,043 true positives for class 0 and 4,956 false negatives, alongside 4,927 true positives for class 1 and 5,075 false negatives. The Logistic Regression model, however, exhibits a slightly different distribution, correctly identifying 5,323 instances of class 0 and misclassifying 4,675 as class 1, while for class 1, it accurately predicted 5,296 instances and misclassified 4,707. Overall, these matrices reveal the strengths and weaknesses of each model, with the Random Forest demonstrating a higher true positive rate for class 0, while KNN and Logistic Regression show more balanced but still suboptimal performance, indicating areas for improvement in classification accuracy across all models.

Customer Segmentation Trends

In the customer segmentation area, the targeting of the high-value customer base is of the highest priority to guide targeted marketing strategies and enhance overall company performance. With the aid of clustering algorithms such as K-Means and Hierarchical clustering, the customer base was segmented into groups based on several attributes such as purchase frequency, average purchase value, and customer lifetime value (CLV). From this segmentation analysis, some high-value segments with regular purchase behavior and

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

deep spending patterns were revealed. For instance, a segment comprised of customers with regular high-value purchase behavior demonstrated a strong propensity toward premium products. With the targeting of the high-value customer base, companies can align their marketing strategies, such as personalized offers or special offers, to enhance customer loyalty and enhance repeat purchase behavior.

Additionally, analyzing the behavior of the segments unveiled preferences and patterns that can guide the development of specific marketing strategies. For example, the value segment of the customer can be targeted with rewards of loyalty or personalized messages that target premium offers. However, the lower-value segments can benefit from educational information or incentives that target increased engagement levels of spending. With the support of the information provided by the analysis of the behavior of the segments, the companies can tailor their marketing strategies to best leverage their assets to maximize customer retention while increasing overall revenues.

Fraud Detection Analysis

Fraud detection is of paramount importance to the protection of financial transactions and customer trust. In this analysis, we compared the performance of a range of machine learning models in detecting fraudulent transactions by their recall rates and precision levels. Using algorithms like Random Forest and Logistic Regression, we were able to successfully pick out a considerable number of fraudulent transactions, with Random Forest standing out with a very strong performance at detecting actual positive results. The models were all trained on a record of past transactions to learn to recognize patterns of behavior that are associated with fraud, like spending patterns that are out of the norm or payments made at unfamiliar locations.

However, while detecting fraud is essential, it is equally important to assess the models' effectiveness in minimizing false positives, as excessive false positives can lead to customer dissatisfaction and financial losses. Our evaluation revealed that the Random Forest model not only excelled in identifying fraudulent transactions but also maintained a lower false positive rate compared to other models. This balance is crucial, as it reduces the number of legitimate transactions flagged as fraudulent, thereby preventing unnecessary disruptions to customer experience. By continuously monitoring and refining these models, businesses can enhance their fraud detection capabilities, ultimately leading to reduced financial losses and a more secure transaction environment for customers.

Practical Implications

Impact on E-Commerce Business Strategies

Machine learning-driven insights have transformed e-commerce business strategies by enabling data-driven strategies to improve pricing, marketing, and stock management. Using algorithms that consider historical sales patterns, rival pricing, and market conditions, companies can implement dynamic pricing strategies that change according to real-time conditions to maximize profit margins while being cost-effective. For example, machine learning models can recognize patterns of customer behavior to forecast patterns of change in customer needs and prices accordingly. In this way, products are sold at the best possible price to maximize both the volume of sales and the revenue generated.

Moreover, marketing is also supported significantly by machine learning by enabling customer segmentation-driven campaigns and behavior analysis-driven campaigns. AI-driven insights are the key to personalizing marketing communications to the customer, with the customer being shown offers that are relevant to them. Not only is this personalization likely to improve the conversion rate, but customer loyalty is also enhanced with the customer feeling valued by being addressed with their unique needs being addressed to them. In addition to this, stock management is also simplified with the support of predictive analysis that forecasts the demand and adjusts the stock accordingly. With the ability to provide best-seller stock without unnecessary stock levels being maintained, the cost is reduced while operational effectiveness is maximized. Overall, the presence of machine learning within e-commerce strategies maximizes customer engagement, generates sales, and maximizes the usage of resources.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Fraud Prevention and Risk Management

Integrating machine learning models with real-time anti-fraud systems is a key innovation in anti-fraud measures and risk management. Using complex algorithms that examine the data of transactions in real-time enables companies to detect potentially fraudulent activity while the activity is taking place. Machine learning models learn to pick up patterns of activity that are associated with fraud, like atypical amounts of a transaction or geo-inconsistencies. With this anticipatory measure, companies can suspend suspicious transactions the moment they are made, initiating additional verification procedures before their final processing.

The early identification of fraud is the key to reducing financial loss. Organizations can avoid financial loss by addressing the potential of fraud at the transactional level while maintaining customer trust intact. The models can also learn to improve with the introduction of new information over time, making them increasingly effective with time. With this capacity to learn, the systems can keep up with the change of approach by the perpetrators of the fraud. With the inclusion of intricate models into their systems, the companies not only secure their assets but also provide a secure environment to their customers while solidifying their image of being transparent and reliable.

Scalability and Future Applications

The scalability of machine learning applications extends beyond e-commerce, offering substantial benefits to a variety of industries such as retail, fintech, and digital marketing. The methodologies developed for customer segmentation, pricing optimization, and fraud detection can be effectively applied to these sectors, enabling organizations to harness the power of data-driven decision-making. For instance, in retail, machine learning can enhance supply chain management by predicting demand trends and optimizing logistics, ensuring that products are delivered efficiently and cost-effectively.

In fintech, tracking of real-time customer behavior can be applied to enhance customer experience and tailor financial services to individuals. Machine learning algorithms can examine spending patterns to provide personalized financial advice and recommend products that are aligned with specific customer needs. AI-powered recommendation engines can also revolutionize digital marketing by providing personalized content that is engaging to the user base, boosting engagement and conversions. As businesses increasingly adopt the technologies of machine learning, the possibilities of applications are immense. With real-time analysis, innovations can occur across various domains, enabling companies to change quickly to adapt to the latest market conditions and customer needs. With the aid of machine learning, companies can not only improve operational effectiveness but also deliver a personalized and compelling customer experience that sets them up to succeed in the long term amid a highly data-driven environment.

Discussion and Future Directions

Challenges of AI-Powered E-commerce Analytics

The implementation of AI-powered analytics within e-commerce is plagued with challenges, especially with respect to dealing with skewed datasets to detect fraud. Fraudulent activity is usually a small percentage of overall activity, causing severe class skew that can adversely impact the performance of models. Conventional machine learning algorithms can fail to properly discover fraudulent activity if they are learned with skewed datasets like this, tending to produce very high rates of false negatives. To overcome this, oversampling the minority class, under-sampling the majority class, or the usage of sophisticated techniques like anomaly detection are necessary measures. Yet they can add their own sets of complications and have the potential to undermine the robustness of the model and its ability to generalize.

Another significant problem is the management of biases within customer segmentation and personalization. Machine learning algorithms have the potential to reinforce biases within the training data unintentionally, leading to skewed results that can positively impact certain customer groups at the expense of others. For instance, if the training data itself is biased by historically skewed purchasing patterns, the

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

resultant segmentation can reinforce stereotypes with less optimal marketing campaigns being the results. Not only is this ineffective at personalization attempts, but also potentially alienates customer groups. To circumvent this sort of bias, fairness-aware algorithms must be put into practice with regular checking of the output of models to treat all customer groups fairly.

Limitations of the Study

While this work is enlightening regarding the application of machine learning to e-commerce analysis, observe that this is also bounded by the limitation of the external validity of synthetic data to real e-commerce activity. It can work well to initialize models and carry out testing with synthetic data; however, the entire variability and sophistication of real customer behavior could fail to be represented by synthetic sets of data. Real activity is influenced by a myriad of external factors such as economic conditions, seasonality patterns, and customer sentiment that synthetic sets of data could fail to adequately simulate. Therefore, the results could fail to fully transfer to live environments.

Moreover, deploying models into operational environments also comes with challenges of its own. Model drift, where the performance of the model degrades with time due to evolving distributions of the data, can have a major impact on effectiveness. Operationalising AI systems also needs strong infrastructure, round-the-clock monitoring, and regular updates to accommodate emerging patterns of customer behavior or types of fraud attempts. All of this creates the need to have a complete plan of maintaining and evolving the models to provide continuous performance and trustworthiness.

Future Research Prospects

Future research into AI-driven e-commerce analysis is rich with opportunity, with the exploration of deep learning techniques to enhance customer behavior models being a key area of research potential. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs), two deep learning models, have shown the potential to learn intricate patterns within large datasets that can lead to more precise customer behavior and preference predictions. With the potential of deep learning at their fingertips, experts can develop increasingly intricate models that develop a deeper knowledge of customer journeys and interactions that ultimately lead to increased personalization and marketing effectiveness.

Furthermore, integrating real-time AI-driven recommendation systems with machine learning is another potential area of research opportunity. With the dynamic change of customer preferences, real-time analysis can enhance the timeliness of recommendation systems to deliver the best possible recommendation at the time of purchase decision-making to the customer. The two can revolutionize the shopper experience by boosting engagement and conversions by integrating the two. It will be imperative to research the potential of the intersection of real-time processing of information with machine learning algorithms to build systems that learn to adapt to customer behavior as well as look ahead to forecast needs. With the research opportunities being pursued by the industry, the industry can go on to build innovative solutions that enhance the effectiveness of e-commerce strategies.

Conclusion

The primary purpose of this research was to leverage the potential of machine learning to extract meaningful information out of synthetic e-commerce data to overcome the limitations of traditional analysis. This research focused on the consumption patterns and trends of the US e-commerce market with specific knowledge of its idiosyncrasies and challenges. The synthetic e-commerce dataset comprises a comprehensive collection of simulated transactional data designed to reflect the dynamics of an online retail environment. The dataset included detailed records of customer transactions, capturing essential information such as transaction IDs, timestamps, product categories, quantities purchased, and total transaction values. Additionally, customer demographics are represented, encompassing attributes such as age, gender, location, and income levels, which facilitate deeper insights into consumer behavior and preferences. The dataset also featured product categories that range from electronics to apparel, allowing for diverse analyses of purchasing trends across different market segments. For model choice, we applied

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

various machine learning algorithms specific to our needs of predicting sales, customer segmentation, demand forecast, and fraud detection. Random Forest, Logistic Regression, and K-Neighbors Classifier are the models selected by the analyst. For evaluating the performance of the models, we used a suite of metrics specific to the task at hand. For the case of fraud detection, the metrics included were accuracy, precision, recall, F1-score, and ROC-AUC. The heights of the bars indicated the accuracies of the models, with Random Forest being the highest, followed by KNN, and Logistic Regression being the least among the two models. The varying heights of the bars pictorially display the comparative performance of the models, with Random Forest being well ahead of the other two models. Machine learning-driven insights have transformed e-commerce business strategies by enabling data-driven strategies to improve pricing, marketing, and stock management. Using algorithms that consider historical sales patterns, rival pricing, and market conditions, companies can implement dynamic pricing strategies that change according to real-time conditions to maximize profit margins while being cost-effective. Integrating machine learning models with real-time anti-fraud systems is a key innovation in anti-fraud measures and risk management. Using complex algorithms that examine the data of transactions in real-time enables companies to detect potentially fraudulent activity while the activity is taking place.

References

Agnes, A. G., Su, H. K., & Kuo, W. K. (2024, June). Personalized E-commerce: Enhancing Customer Experience through Machine Learning-driven Personalization. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-5). IEEE.

Akter, R., Nasiruddin, M., Anonna, F. R., Mohaimin, M. R., Nayeem, M. B., Ahmed, A., & Alam, S. (2023). Optimizing Online Sales Strategies in the USA Using Machine Learning: Insights from Consumer Behavior. Journal of Business and Management Studies, 5(4).

Al Montaser, M. A., Ghosh, B. P., Barua, A., Karim, F., Das, B. C., Shawon, R. E. R., & Chowdhury, M. S. R. (2025). Sentiment analysis of social media data: Business insights and consumer behavior trends in the USA. Edelweiss Applied Science and Technology, 9(1), 545-565.

Buiya, M. R., Laskar, A. N., Islam, M. R., Sawalmeh, S. K. S., Roy, M. S. R. C., Roy, R. E. R. S., & Sumsuzoha, M. (2024). Detecting IoT Cyberattacks: Advanced Machine Learning Models for Enhanced Security in Network Traffic. Journal of Computer Science and Technology Studies, 6(4), 142-152.

Chandel, A. (2024). Analytics: Leveraging Real-Time Data. Improving Entrepreneurial Processes Through Advanced AI, 267.

Chopra, N., Patel, A., Singh, N., & Sharma, V. (2020). Leveraging Reinforcement Learning and Neural Networks for Optimized Dynamic Pricing Strategies in E-Commerce. International Journal of AI Advancements, 9(4).

Hasan, M. R., Islam, M. R., & Rahman, M. A. (2025). Developing and implementing AI-driven models for demand forecasting in US supply chains: A comprehensive approach to enhancing predictive accuracy. Edelweiss Applied Science and Technology, 9(1), 1045-1068.

Islam, M. R., Nasiruddin, M., Karmakar, M., Akter, R., Khan, M. T., Sayeed, A. A., & Amin, A. (2024). Leveraging Advanced Machine Learning Algorithms for Enhanced Cyberattack Detection on US Business Networks. Journal of Business and Management Studies, 6(5), 213-224.

Islam, M. Z., Islam, M. S., Reza, S. A., Bhowmik, P. K., Bishnu, K. K., Rahman, M. S., ... & Pant, L. (2025). Machine Learning-Based Detection and Analysis of Suspicious Activities in Bitcoin Wallet Transactions in the USA. Journal of Ecohumanism, 4(1), 3714-3734.

Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (2022). Optimizing E-Commerce Revenue: Leveraging Reinforcement Learning and Neural Networks for AI-Powered Dynamic Pricing. International Journal of AI and ML, 3(9).

Kasemrat, R., & Kraiwanit, T. (2023). Benchmarking Machine Learning Models for Predictive Analytics in E-Commerce. Available at SSRN 4832967.

Mohaimin, M. R., Das, B. C., Akter, R., Anonna, F. R., Hasanuzzaman, M., Chowdhury, B. R., & Alam, S. (2025). Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market. Journal of Computer Science and Technology Studies, 7(1), 30-45.

Nabi, N., Pabel, M. A. H., Rahman, M. A., Mozumder, M. A. S., Al-Imran, M., Sweet, M. M. R., ... & Sharif, M. K. (2024). Unleashing Deep Learning: Transforming E-commerce Profit Prediction with CNNs. Journal of Business and Management Studies, 6(2), 126-131.

Rana, M. S., Chouksey, A., Das, B. C., Reza, S. A., Chowdhury, M. S. R., Sizan, M. M. H., & Shawon, R. E. R. (2023). Evaluating the Effectiveness of Different Machine Learning Models in Predicting Customer Churn in the USA. Journal of Business and Management Studies, 5(5), 267-281.

Rahman, A., Debnath, P., Ahmed, A., Dalim, H. M., Karmakar, M., Sumon, M. F. I., & Khan, M. A. (2024). Machine learning and network analysis for financial crime detection: Mapping and identifying illicit transaction patterns in global black money transactions. Gulf Journal of Advance Business Research, 2(6), 250-272.

Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing Fraud Detection Systems in the USA: A Machine Learning Approach to Identifying Anomalous Transactions. Journal of Economics, Finance and Accounting Studies, 5(5), 145-160.

Journal of Ecohumanism
2025
Volume: 4, No: 2, pp. 2394 – 2420
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i2.6635

Ray, R. K., Sumsuzoha, M., Faisal, M. H., Chowdhury, S. S., Rahman, Z., Hossain, E., ... & Rahman, M. S. (2025). Harnessing Machine Learning and AI to Analyze the Impact of Digital Finance on Urban Economic Resilience in the USA. Journal of Ecohumanism, 4(2), 1417-1442.

Saleem, H., Muhammad, K. B., Nizamani, A. H., Saleem, S., & Aslam, A. M. (2019). Data science and machine learning approach to improve E-commerce sales performance on social web. International Journal of Computer Science and Network Security (IJCSNS), 19.

Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

Sizan, M. M. H., Chouksey, A., Miah, M. N. I., Pant, L., Ridoy, M. H., Sayeed, A. A., & Khan, M. T. (2025). Bankruptcy Prediction for US Businesses: Leveraging Machine Learning for Financial Stability. Journal of Business and Management Studies, 7(1), 01-14.

Sizan, M. M. H., Chouksey, A., Tannier, N. R., Al Jobaer, M. A., Akter, J., Roy, A., ... & Islam, D. A. (2025). Advanced Machine Learning Approaches for Credit Card Fraud Detection in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 883-905.

Sumsuzoha, M., Rana, M. S., Islam, M. S., Rahman, M. K., Karmakar, M., Hossain, M. S., & Shawon, R. E. R. (2024). LEVERAGING MACHINE LEARNING FOR RESOURCE OPTIMIZATION IN USA DATA CENTERS: A FOCUS ON INCOMPLETE DATA AND BUSINESS DEVELOPMENT. The American Journal of Engineering and Technology, 6(12), 119-140.

Tandon, U., Tandon, A., & Mehrotra, T. (2024). Transformation in the World of Commerce and Economics through AI. In Artificial Intelligence: A Multidisciplinary Approach towards Teaching and Learning (pp. 194-215). Bentham Science Publishers.

Ye, X., & Jonilo, M. (2023). Unleashing the Power of Big Data: Designing a Robust Business Intelligence Framework for E-commerce Data Analytics. Journal of Information Systems Engineering and Management, 8(3), 22638.

Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G., & Wu, J. (2023). A brief survey of machine learning and deep learning techniques for e-commerce research. Journal of Theoretical and Applied Electronic Commerce Research, 18(4), 2188-2216.