CE477 - Project

Yankı Omaç
Electrical and Electronics Engineering
İzmir University of Economics
İzmir, Turkey
Email: yanki.omac@std.izmirekonomi.edu.tr

Recep Atabey Demir
Computer Engineering
İzmir University of Economics
İzmir, Turkey
Email: atabey.demir@std.izmirekonomi.edu.tr

Bersay Yakıcı
Software Engineering
İzmir University of Economics
İzmir, Turkey
Email: bersay.yakici@std.ieu.edu.tr

Halil Arda Aşılıoğlu
Software Engineering
İzmir University of Economics
İzmir, Turkey
Email: arda.asilioglu@std.izmirekonomi.edu.tr

***Abstract*—**

An era where e-commerce is booming, the ability to understand and optimize customer experience is paramount for businesses aiming to thrive. An international e-commerce company, specializing in electronic products, has embarked on an ambitious project to delve deep into their customer database to uncover vital insights that could revolutionize their operations. Leveraging advanced machine learning techniques, the company aims to dissect the complex dynamics of customer interactions and product shipments to enhance satisfaction and efficiency.

I. Introduction

In today's thriving e-commerce landscape, understanding and enhancing the customer experience are essential for business success; an international e-commerce company specializing in electronic products is embarking on a data-driven journey to revolutionize its operations, leveraging advanced machine learning techniques to dissect customer interactions and product shipments, aiming to boost satisfaction and efficiency with a foundation of a robust dataset of 10,999 observations across 12 variables providing insights into the entire customer journey from purchase to delivery. This project aims to equip the enterprise with insights to surpass customer expectations and improve operational efficiency through data-driven strategies. E-commerce research emphasizes the importance of data-driven approaches in customer segmentation, recommendation systems, and supply chain management. Existing literature and works on similar datasets highlight the transformative potential of machine learning in extracting insights from e-commerce datasets. The importance of delivering the goods on time by customer reviews and trying to predict ways to deliver data in faster ways. The goals of the project include identifying patterns in the customer journey, segmenting customers for targeted strategies, optimizing supply chain operations, and enhancing overall customer satisfaction. This project seeks to empower the enterprise with analytical tools for navigating the e-commerce landscape and fostering customer-centric growth. [1]

II. Data Set Description

The foundation of this analytical venture is a robust dataset comprising 10,999 observations across 12 meticulously curated variables. These variables provide a comprehensive overview of the customer journey, from the initial purchase to the final delivery. Data set is a clean data set and it has no null or missing values.

ID: A unique identifier for each customer, ensuring precise tracking and personalized insights.

Warehouse Block: With the company's expansive warehouse segmented into blocks A through E, this variable helps in logistics optimization and inventory management.

Mode of Shipment: Understanding the impact of different shipment methods (Ship, Flight, Road) on customer satisfaction and delivery efficiency.

Customer Care Calls: The frequency of customer inquiries serves as an indicator of service quality and customer engagement.

Customer Rating: A direct measure of customer satisfaction, with ratings ranging from 1 (lowest) to 5 (highest).

Cost of the Product: This financial metric is crucial for pricing strategies and profitability analysis.

Prior Purchases: Tracking customers' purchase history aids in predicting future buying behavior and personalizing marketing efforts.

Product Importance: Categorizing products based on their importance (low, medium, high) enables tailored handling and prioritization.

Gender: Analyzing shopping patterns and preferences across genders.

Discount Offered: Examining the impact of discounts on sales volume and customer acquisition.

Weight in Grams: The logistical aspect of shipping, influencing costs and delivery methods.

Reached on Time: The critical outcome variable indicating whether a product was delivered within the expected timeframe, serving as a benchmark for operational efficiency. 0 indicates it reached on time and 1 indicates that its not reached on time.

III. PREPROCESSING

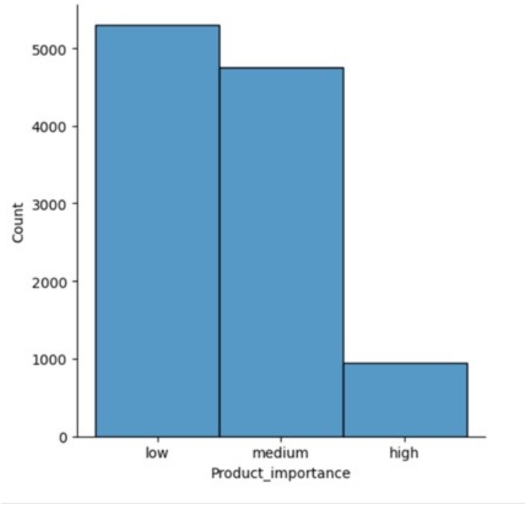

Fig. 1. Histogram of Product Importance

Figure 1, illustrates the distribution of product importance across different categories. The counts for each category are as follows:

Low Importance: Over 5000 occurrences. Medium Importance: Between 4500 and 5000 occurrences. High Importance: Between 750 and 1000 occurrences. The analysis suggests that a significant proportion of products fall into the low importance category, while fewer products are classified as high importance.

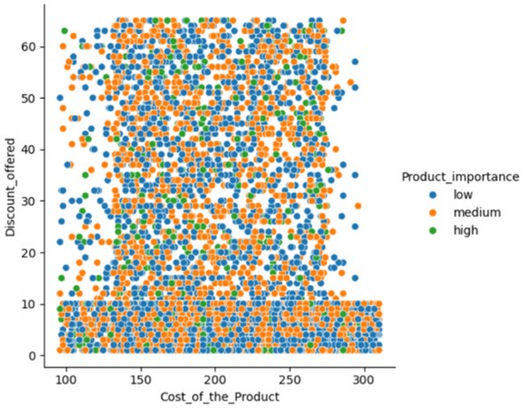

Fig. 2. Scatter plot of Cost and Discount

Figure 2, illustrates the relationship between the cost of a product and the discount offered. Each data point represents a product, color-coded by its importance level (low, medium, or high).

X-axis: Represents the cost of the product (ranging from 0 to 300). Y-axis: Represents the discount offered (ranging from 0 to 60). Observations:

Data points are widely scattered, indicating variability in discounts across different product costs and importance levels. No clear linear trend is visible; the distribution is diverse. The dense concentration of points occurs across all levels of product importance, especially in the mid-range of cost and discount. This scatter plot provides insights into the complex relationship between cost and discount, suggesting that discounts are not consistently tied to either product cost or importance.

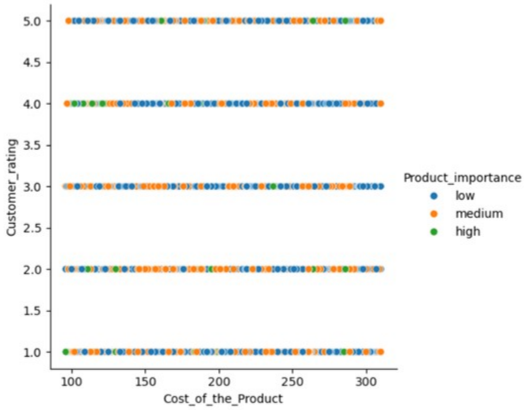

Fig. 3. Scatter plot of Cost and Customer

Figure 3, represents customer ratings in relation to the cost of the product. The x-axis displays the cost of the product, ranging from 100 to 300. The y-axis represents customer ratings, which vary from 1.0 to 5.0. Regardless of the cost and importance level, products consistently receive a wide range of ratings. Data points are distributed evenly across all levels of customer ratings.

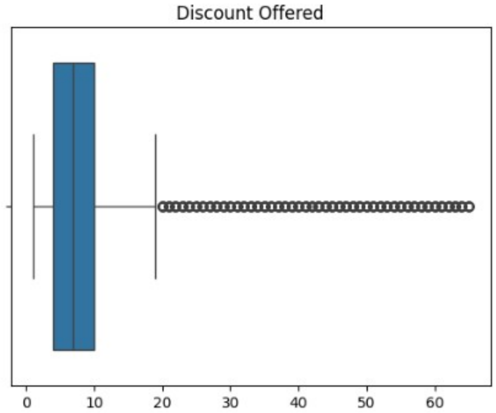

Fig. 4. Box plot of Discount

Figure 4, illustrates the discounts provided by a certain entity. The x-axis represents the percentage of discount, ranging from 0 to 60. Notably, a prominent blue bar extends upwards around the 10 mark on the x-axis. This suggests that a significant number of offers fall within this range. The presence of error bars on both sides of the blue bar indicates variability or a range in which these discounts are typically offered.There are no lower outliers, indicating that extremely low discounts are not common. However, the upper end of the data exhibits a cluster of outliers. These outliers extend beyond the IQR, reaching up to nearly 60 discount. While such high discounts are infrequent, they do exist.

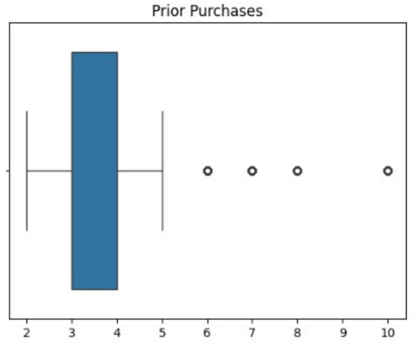

Fig. 5. Box plot of Prior Purchases

The blue rectangular box in the Figure 5 represents the interquartile range (IQR). This range encapsulates the middle half of the observed prior purchase values. Specifically, the lower edge of the box corresponds to approximately 3 on the x-axis, while the upper edge aligns with around 4. This implies that half of the customers made prior purchases within this range. The line inside the blue box represents the median value. In this case, the median is close to 3 on the x-axis. The median serves as a robust measure of central tendency, indicating that half of the customers made three or fewer prior purchases. Beyond the upper whisker of the box, we observe several outliers represented by circles. These outliers correspond to individual observations that deviate significantly from the bulk of the data. Specifically, these customers have made notably more purchases, ranging from approximately 6 to nearly 10.

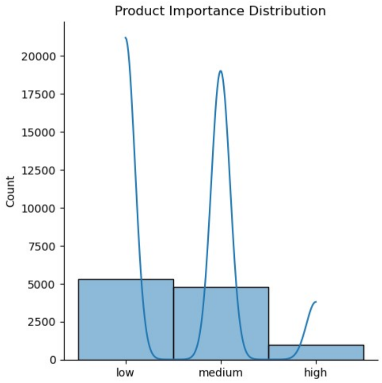

Fig. 6. Histogram of Product Importance

Figure 6, represents the frequency distribution of product importance levels. Three importance categories are considered: low, medium, and high. The y-axis represents the count of products falling into each importance category. The count ranges from 0 to approximately 20,000. Medium Importance: The most prevalent category, with a peak count near 20,000. Low Importance: Follows medium importance, indicating a substantial but lesser count. High Importance: Notably less common, with a small spike in its distribution. The concentration of products lies in the medium importance range, suggesting a balanced inventory or product offering. Moderately important items dominate, while highly essential and less critical products are relatively infrequent.

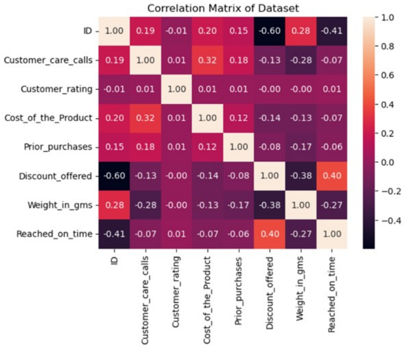

Fig. 7. Heatmap of Numerical data

Figure 7, illustrates the correlation coefficients between various numerical variables. The color gradient on the right side of the heatmap corresponds to the range of correlation coefficients. It spans from -0.4 (depicted in dark purple) to 1.0. Warmer colors (reds and oranges) indicate positive correlations, while cooler colors (purples) represent negative correlations. Each cell in the matrix displays a numerical value representing the correlation coefficient between a pair of variables. Notably, there is a positive correlation of 0.40 between Discount offered and Reached on time. This suggests that as the discount offered increases, the likelihood of orders being delivered on time also rises. Additionally, Customer care calls and Cost of the Product exhibit a positive correlation of 0.32. Conversely, there is a negative correlation of -0.38 between Discount offered and Weight in gms. This implies that higher discounts may be associated with lighter products.

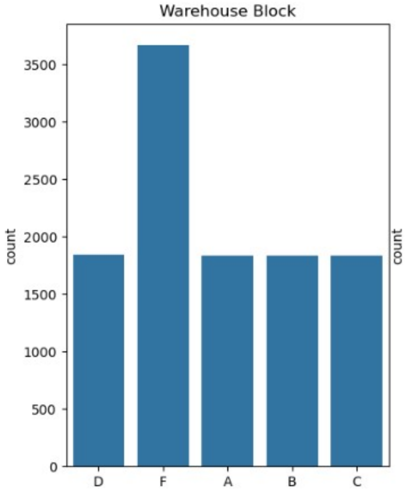

Fig. 8. Distribution of Warehouses

Figure 8, illustrates the distribution of warehouses across different blocks labeled D, F, A, B, and C. Block F has the highest count of warehouses, exceeding 3500. It represents the most densely populated area in terms of warehouses. Other blocks have similar counts of warehouses, both slightly below 2000. While they don't have as many warehouses as Block F, they still contribute significantly to the overall distribution.

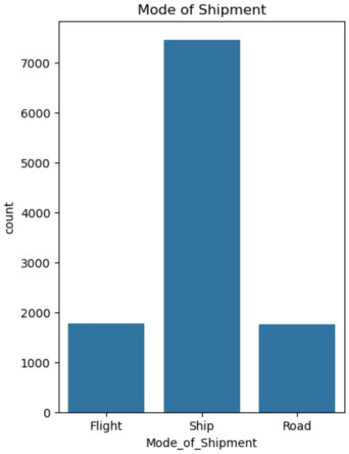

Fig. 9. Distribution of Shipment

Figure 9, illustrates the distribution of shipments across three primary modes of transport: flight, ship, and road. Each mode is represented by a distinct blue bar. Notably, shipping emerges as the dominant choice, accounting for over 7,000 shipments. In contrast, both flight and road transport exhibit significantly lower counts, hovering just above 1,000 each. This data underscores the reliance on maritime transport within the context of our study.

Z-Score Standardization: Sets the mean of each feature to 0 and the standard deviation to 1. Subtracts each value from the feature's mean and divides it by the standard deviation. As a result of this process, the distribution of the data resembles a normal distribution with mean 0 and standard deviation 1. In this way, the effect of outliers is reduced and model performance increases.

Min-Max Normalization : Scales values within a certain range. Rescales the values for each feature so that the minimum value is 0 and the maximum value is 1. In this way, all values are between 0 and 1.

PCA: PCA is a dimensionality reduction technique that is used for attribute selection. We used this technique to find out the most important columns in our data set. The most important 2 columns are Customer Care Calls and Discount offered columns which has highest variance ratios over other columns.

Discretization: Discretization process was applied on numerical attributes. Specific ranges were determined for the `Cost_of_the_Product` and `Weight_in_gms` attributes and were categorized according to these ranges.

For `Cost_of_the_Product`, the ranges [0-100], [100-200], [200-300], [300-400], [400-500] are determined and 'Very Low', 'Low', 'Medium', 'High', 'Very High' categories.

For `Weight_in_gms`, the ranges [0-1000], [1000-2000], [2000-3000], [3000-4000], [4000-5000] are determined and 'Very Light', 'Light', 'Moderate', 'Heavy', 'Very Heavy' categories.

As a result of these operations, the data set was updated according to the ranges and categories determined for both attributes.

References

[1] W. O. Gibin, "On-time delivery dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/willianoliveiragibin/on-time-delivery/code